Beyond Recall: Behavioral Specification as an Interpretive Layer for AI Personalization

Aarik Gulaya*

April 2026

Abstract

Abstract to be written.

Author: Aarik Gulaya, Base Layer (`aarik@base-layer.ai`, `base-layer.ai`) **Date:** April 2026 **Preprint** (manuscript CC-BY-4.0; code Apache 2.0) **Data + Code:** github.com/agulaya24/base-layer **Study Repository:** github.com/agulaya24/beyond-recall

---

1. Introduction

1.1 Recall is not interpretation. Interpretation can be measured.

AI is moving from a tool a person uses to an agent that acts on a person's behalf, and that shift changes what "memory" must do for a specific individual. State of the art AI memory has been optimizing for recall as the success metric. The four leading systems (Zep, Letta, Mem0, and Supermemory) compete on standard recall benchmarks such as LOCOMO and LongMemEval, reporting accuracies in roughly the 70% to 93% range depending on provider, model, and benchmark variant (§2.2). Optimizing further on recall leaves something more fundamental unmeasured. This research paper explores how recall is one part of memory, and how the function of memory is dictated by how an individual processes the facts and experiences of their life.

We use **interpretation** to refer to this human-side property: the way a specific person processes facts and experiences into judgments, decisions, and reactions. Viewing situations from different lenses can lead to entirely different interpretations of the same set of facts. This has been shown across the human experience, from the sciences to religion, and by extension to the relative experiences of any individual; memory is deeply personal. For an AI to serve a specific person, it must be given context on the framework that person uses to reason, not just the raw facts or information itself. Throughout this paper we use the term **Behavioral Specification** to refer to a static document that extracts and encodes a person's behavioral patterns; the operational definition is developed across §3.7. The Behavioral Specification is the artifact that captures this interpretive framework, and is provided to an AI as context.

We introduce **representational accuracy** as the corresponding AI-side property: how well a system's internal model of a specific person captures their interpretive patterns. It is not recall,

*aarik@base-layer.ai; independent researcher.

preference matching, or persona consistency. It is a distinct property of the AI system, and state of the art memory benchmarks do not isolate it. Prior work closest to this axis (Twin-2K for scaled behavioral prediction, PersonaGym for persona fidelity, AlpsBench for preference alignment) measures related properties but not the transfer of a person's interpretive patterns to new situations the system has never seen. §2.1 positions each benchmark against what this paper measures, and Appendix F develops the scope differences in detail.

The core hypothesis of this research is that representational accuracy of a person's interpretation improves an AI system's behavioral alignment with that person. This is the operational primitive for any AI system meant to act on a person's behalf: the system's behavior can only match the user's reasoning to the extent the system represents that reasoning accurately. The operational test in this paper is behavioral prediction on held-out situations: given a situation drawn from text the model has never seen, the model generates how the subject would respond; the response is scored by a panel of calibrated large language model (LLM) judges against the subject's own verbatim response in the held-out text on a 1-5 interpretive rubric (§3.3). Accurate prediction on held-out text is evidence that the representation captures the subject's recurring patterns of reasoning, distinct from the facts and stylistic surface that current extraction pipelines already produce. The design also reduces the risk of sycophancy[1]: the answer is checked against the person's narrative, which the model has never seen, not against anything the user says during the conversation. The held-out test is one operationalization of the hypothesis.

We test this hypothesis on the leading state-of-the-art AI memory systems and on a diverse set of 14 autobiographies from authors across the world. For this initial examination we use baselined and calibrated LLM judges to evaluate the performance of each memory system, on its own and in combination with a **Behavioral Specification**: a static document that extracts and encodes a stable representation of a corpus's behavioral patterns. The specification captures the recurring patterns in how the subject reasons, drawn from the shape of judgments and reactions across the corpus (for example: *"spiritual integrity over social cost..."*, *"reform through love..."*, *"hierarchical deference..."*). A walked example of the audit chain from such a pattern back to its grounding facts and source passages appears in §2.3.

Defined terms used throughout the paper are collected in **Appendix H** for reference.

1.2 What we tested

We tested the Behavioral Specification across 14 historical subjects, each with a public domain autobiography. For every subject we split the source corpus in half: the training half was used to generate the specification, to seed each memory system, and to provide the retrievable fact pool. The held-out half was used only to produce behavioral prediction questions and was never shown to the **response model**, the language model being asked to predict how the subject would respond. The Behavioral Specification is the context document that the response model receives. The set of held-out questions for each subject is the **question battery** (size and composition per subject in §3.5). The test was whether each system, under each tested condition, could predict how that specific person would respond in situations drawn from text it had never seen.

[1]Sycophancy refers to a model adjusting its answer to match what the user appears to want, often by agreeing with or flattering them. See Sharma et al. (2023), Perez et al. (2022), and §2.4 (Jain et al., 2025). Whether an accurate representation also produces aligned behavior on situations the person never encountered, and what else such a representation enables (steering, agent action on the person's behalf), are open questions of the broader research program (§7).

The Behavioral Specification itself is built from the training-half corpus through an extraction-and-authoring pipeline (§3.7). The pipeline distills the recurring patterns of how the subject reasons into a single structured document, typically around 7,000 tokens (~5,000 words) long. That document is what the response model receives as context when asked to predict how the subject would respond.

Hypotheses. The study tests five claims about how a representation of a person shapes AI behavior on that person's behalf:

- **H1.** A response model given a Behavioral Specification produces responses that align with the person's documented behavior more closely than the same model given no context, facts retrieved by a memory system, the full extracted fact list, or the raw source corpus (§4.1).
- **H2.** The specification's benefit is inversely proportional to the response model's pretraining coverage of the person. Its effect is largest on people the model does not already know (§4.1).
- **H3.** The benefit comes from the content of the correct specification for the correct person, not from the mere presence of a structured prompt. A random other person's specification, applied in its place, produces a substantially smaller and content-specific effect than the matched specification: adversarial pairings degrade performance below the no-context baseline, while random derangements produce only a small residual effect (§4.3).
- **H4.** The specification interacts with memory-system retrieval in a structured way that depends on the type of question being asked. Aggregate effects on each memory system reflect the balance of these per-question patterns and shift with retrieval architecture (§4.4).
- **H5.** The Behavioral Specification's quality advantage is also a compression advantage: a ~7,000-token (~5,000-word) specification recovers most of the predictive accuracy of an 80-400K-token (~60-300K-word) raw corpus (§4.2).

Post-hoc analyses surfaced during the work are reported alongside these results.[2]

Primary and secondary outcomes. The **primary outcome** is the mean prediction score on the 1-5 rubric across a 5-judge primary panel (§3.3).[3] Cross-subject claims are calculated subject-by-subject before averaging, so they are not driven by subjects with larger question batteries. As a **secondary outcome**, we report the per-question **improvement rate**: how often a context condition helps relative to the comparison baseline (§4.2.1), not just by how much it helps when averaged. The per-question secondary outcome is informative because the Spec's effect is conditional on question type: aggregate effects reflect the balance of interpretation-heavy items (where the Spec lifts most) and literal-recall items (where retrieval already suffices). The formal proposal and failure-mode analysis for the secondary outcome are in §4.2.1; full operational details for both outcomes are in §3.3.

Each memory system is tested in both a controlled configuration (identical pre-extracted fact pool) and a native configuration (the provider's own ingestion pipeline); design detail in §3.2. Running in parallel across both is the Behavioral Specification, tested alone and layered on top of each configuration. Every meaningful combination of inputs is evaluated as its own condition:

[2]Several post-hoc analyses surfaced during the work and are reported alongside the H1–H5 results: the cross-system retrieval-overlap divergence (§4.4.1, with sensitivity in §4.6.6), the Letta stateful-agent case study (§4.5; Appendix G), the abstention-credit validity audit (§3.3.6), and the per-subject wrong-Spec heterogeneity table (§4.6.5). These are labeled where they appear and are reported as exploratory rather than at the same evidentiary tier as the pre-registered hypotheses. Full breakdown in Appendix B.10.

[3]Aggregation rule (the "locked rule" referenced throughout): per-question scores within a (subject, condition) cell are first averaged across the 5 judges, then per-subject means are aggregated across the 14 subjects. Full mechanics in §3.3.5.

Group	Condition	Inputs given to the model	Purpose
Direct context manipulations	**No context** (C5)	Nothing. The model answers from pretraining alone.	Pretraining baseline. Measures what the model already knows about the subject from public sources.
	Specification alone (C2a)	The Behavioral Specification, with no retrieval, no facts, and no corpus.	Tests whether structure without retrieval is sufficient on its own.
	Wrong-specification control (C2c)	A different subject's specification applied to this subject. Two variants: an adversarial fixed pairing (v1) and a random derangement (v2).[4]	Tests whether the effect is driven by the content of the correct specification, or by the mere presence of structured prompting.
	All facts, no specification (C4)	Every extracted fact for the subject, loaded into context at once.	Tests whether information sufficiency alone drives prediction, independent of structure.
	Facts + specification (C4a)	Every extracted fact plus the specification.	Combines full information and structure to test the upper bound of context-provided prediction.
	Raw corpus, no specification (C8)	The full training-half corpus loaded into context.	Tests whether unstructured source text can substitute for an interpretive representation.
	Corpus + specification (C9)	Raw training corpus plus the specification.	Tests whether structure is additive to unstructured source text.

[4] v1 is a deterministic fixed pairing that matches each subject with a culturally and temporally distant other (mapping in `scripts/run_global_rerun.py`); v2 applies a random derangement, seed-fixed, so no subject receives its own. Hamerton has an additional variant (Franklin's specification) reported separately in §4.1.2.

Group	Condition	Inputs given to the model	Purpose
Memory-system configurations (controlled, all 5 systems)	**Retrieval alone, controlled** (C1)	Top-k facts retrieved by each memory system (Mem0, Letta[5], Supermemory, Zep, Base Layer) from the shared fact pool.	Tests retrieval sufficiency, and whether providers converge on which facts are most relevant given identical input.
	Retrieval + specification, controlled (C3)	Memory system retrieval from the shared fact pool, plus the specification.	Tests whether the specification layers cleanly on retrieval when the input is held constant.
Memory-system configurations (native, 4 commercial systems)	**Retrieval alone, native** (C1 native)	Top-k results from each memory system's own ingestion pipeline operating over the raw training corpus.	Real-world comparison of each memory system's full ingestion-plus-retrieval stack.
	Retrieval + specification, native (C3 native)	Memory system's own ingestion and retrieval, plus the specification.	Tests whether the specification improves the real-world deployment of each memory system.

The 14 subjects span four continents and roughly two millennia of written human experience. Ordered chronologically: Saint Augustine (North Africa, 4th-5th c.), Bābur (Central Asia and India, 15th-16th c.), Bernal Díaz del Castillo (Spain and Mexico, 15th-16th c.), Benvenuto Cellini (Italy, 16th c.), Jean-Jacques Rousseau (France, 18th c.), Olaudah Equiano (West Africa and Britain, 18th c.), Mary Seacole (Jamaica and Britain, 19th c.), Elizabeth Keckley (United States, 19th c.), Yung Wing (China and the United States, 19th c.), Philip Gilbert Hamerton (Britain, 19th c.), Fukuzawa Yukichi (Japan, 19th c.), Georg Ebers (Germany, 19th c.), Sunity Devee (India, late 19th c.), and Zitkala-Ša (Yankton Dakota, early 20th c.). Source corpora range from 25,231 words (Hamerton) to 422,772 words (Bābur). Full source references are in §3.3.

Predictions were scored on a 1-5 rubric where the integer anchors mark categorical shifts in answer quality (full rubric in §3.3, summarized in the table below). Crossing an integer anchor represents a real change in the kind of answer the model produced, not a small numerical adjustment. For example, a move from 1.8 to 2.4 crosses the 2.0 boundary: the model goes from refusing the question or producing a wholly wrong answer (anchor 1) to engaging with the right topic, even when the specific prediction is still off (anchor 2). Absolute point gains, not percentages, are the informative metric for cross-subject comparison.

[5]Of the four commercial memory systems, Letta is architecturally distinct: alongside retrieval, it maintains a persistent memory block that its agent self-edits during multi-turn conversation. Because this path is not exercised by the retrieval conditions in this table, we ran a separate test on three subjects spanning a 9× corpus-size range (Hamerton, Ebers, Bābur). A fresh Letta agent ingested each training corpus turn-by-turn and was allowed to self-edit. The resulting memory block was then served to the same response model used throughout the main study for a matched comparison against the Behavioral Specification. Full methodology and results in §4.5.

Score	What it means	Shift from previous anchor
1	Refuses or wholly wrong	(rubric floor)
2	Right topic, wrong prediction	From cannot engage to orienting to the question
3	Right domain, no specifics	From wrong prediction to in the neighborhood
4	Right direction with specifics	From in the neighborhood to right direction with specifics
5	Predicts the specific outcome	From right direction to matching the held-out text

Score interpretation, including the cross-anchor rule for fractional scores (e.g., 2.5, 3.4), is in §3.3.1. Example questions per subject and panel composition are in §3.3.2.

The **baseline** we refer to throughout is the no-context condition (C5): the response model's score with no external information. **Low-baseline** subjects are the **population of relevance**: people the model has insignificant pretraining understanding of, even when fragments of their digital footprint exist in training data. **High-baseline** subjects are the opposite, people the model already knows about from pretraining. Almost everyone in the active human population falls into the low-baseline band; even people with substantial public output captured in training corpora have only fragments of their reasoning represented. The low-baseline band is the rule, not the exception.[6] Results are reported separately on the low-baseline slice (n=9) alongside the full 14-subject analysis.

The study is structured into two tiers. Tier 1 (main study) uses Claude Haiku 4.5 as the response model across all 14 subjects on every condition. Tier 2 is a smaller cross-provider directional probe (§3.6, §4.6.1). The 7-judge panel spans three providers; the 5 non-Gemini judges form the primary aggregate and the 2 Gemini judges are reported as a sensitivity check (§3.3.3).

Together these hypotheses test whether a Behavioral Specification can move a language model toward acting in alignment with a specific person.

1.3 What we found

The Behavioral Specification (referred to as the **Spec** in the discussion that follows) acts as an interpretive layer. The Spec's benefit is largest where the model knows the person least, and the mechanism is per-question. On questions where the model needs an interpretive frame and lacks one, the Spec categorically improves the answer produced. On questions where the model already has the answer, the Spec adds nothing and sometimes hurts.[7] What follows are seven findings, beginning with the cross-subject gradient (primary outcome) and the per-question mechanism beneath it. The thread is alignment: how accurately a model predicts a specific person's reasoning is the operational measure of how closely it can act in alignment with that person.

Headline findings.

[6] Operational thresholds: low-baseline C5 $\leq$ 2.0 on the 1-5 rubric, high-baseline C5 $>$ 3.0. Full distribution and band assignments in §3.4.1.

[7] "Low baseline" means C5 $\leq$ 2.0 on the 1-5 rubric. This is the population of importance for AI personalization (§2 lede, §1.4, §5.2): on a frontier model serving general AI users, almost everyone falls in or below this band, even people with substantial public output. The §2 lede defines what "personalization" means in this paper's stronger sense (the interpretive layer beneath stated preferences and biographical facts).

- **Gradient. Every low-baseline subject improved with the Spec; mean lift +0.89 points on the 1-5 rubric, 78.6% of individual questions improve.**[8][9] Across subjects, the Spec's benefit is largest where the model knows the person least. The gradient comes from the per-question mechanism described next: low-baseline subjects have more questions where the model lacks an interpretive frame, and therefore more questions where the Spec lifts. Detail in §4.1.
- **Per-question interpretive lift. The Spec moves 55% of low-baseline questions across at least one rubric anchor upward; 18% cross two or more.** On individual questions where the model needs an interpretive frame and lacks one, the Spec categorically improves the answer produced. Crossing one rubric anchor moves a response from "wrong prediction" to "right direction with specifics." Crossing two or more anchors is a bigger jump: a single question where the model moves from refusal or generic guessing to a recognizable, person-specific response. *5.9% cross three or more anchors.* The pattern holds across Spec-only, facts+Spec, and corpus+Spec conditions on the low-baseline subjects. Detail in §4.1, §4.2.
- **Compression. The Spec recovers 76% of what the corpus delivers at 23× less context.** A 7,000-token structured representation matches most of the predictive accuracy of a 163,000-token raw corpus. The Spec selects and structures the behavioral signal; the interpretive layer drives the result, not the volume of context. *Spec-alone +0.71 vs. corpus-alone +0.93 over baseline. On Hamerton (smallest corpus tested), the Spec scores higher than the raw corpus (2.63 vs. 2.27).* Detail in §4.2.
- **Content specificity. The wrong Spec drops accuracy below the no-context baseline ($\Delta = -0.25$); the correct Spec lifts accuracy above ($\Delta = +0.35$).** What produces the lift is the content of the correct Spec for the correct person, not the presence of a structured prompt. Random pairings (a wrong Spec assigned by chance to a different subject) sometimes still produce predictions that align with the held-out text, suggesting some behavioral patterns transfer across subjects, but the correct Spec consistently outperforms. *Random-derangement $\Delta = +0.15$.* Detail in §4.3.
- **Memory-system layering. The Spec lifts 3 of 4 commercial memory systems on aggregate; per-question anchor crossings range 20–36% across systems.** Layered on top of commercial memory systems, the Spec helps dramatically on interpretation-heavy questions and reduces refusal rates on questions retrieved facts could not ground. It also exposes per-question structure the aggregate hides: some questions improve, some regress, and the balance shifts with retrieval architecture. *Mem0, Letta, and Zep show positive aggregate lift; Supermemory does not.* Detail in §4.4.
- **Hedging reduction. The Spec collapses baseline hedging from 41.2% of responses to 0.4%.**[10] The reduction is content-specific, not prompt-driven: under the wrong-Spec adversarial control, the model continues to hedge or explicitly flag the mismatch on 60.6% of responses (§4.3). Where the Spec content matches the subject, the model commits; where it does not, the model abstains. On low-baseline subjects, where the model would otherwise

[8] The Wilcoxon signed-rank test asks how unlikely it is that the observed pattern of subject-by-subject improvement could happen by chance. Result: W = 11, N = 14, p = 0.007. Full regression (slope −0.96, $R^2 = 0.82$) and the leveler-framing of the gradient (the Spec brings every subject toward roughly the same prediction quality, ~2.44 on the 1-5 rubric per the locked per-subject aggregation rule) are in §4.1.

[9] +0.89 is the cross-subject mean of per-subject Δ_C4a (per-subject grain is the locked unit of inference; see §1.2 aggregation rule). The grand-mean alternative (C4a grand mean minus C5 grand mean) yields +0.93. Detail in Appendix B.9.

[10] Headline number uses the broad rule (any refusal pattern anywhere in the response) under the facts + Spec (C4a) condition. The stricter rule (refusal pattern as the first non-whitespace text) gives 28.8% → 0.0% on the same condition.

refuse to engage, the matched Spec converts refusal into substantive response. This is the gradient operating at its floor. Detail in §4.3.

- **Retrieval divergence. Given identical input, memory-system providers share zero top-10 facts on 35.9% of question-pairs; mean pairwise overlap is 8.3%.** On standard recall benchmarks like LongMemEval and LOCOMO, the four commercial memory systems we tested perform within a few percentage points of each other. Yet on which facts to surface as most relevant, they substantially diverge. Convergence on top-K under identical input would have been evidence of a shared interpretive framework; the systems do not converge. *On 65.6% of (system pair, question) instances they share one or fewer facts.* Detail in §4.4.1.

Mechanism: three patterns of interaction with retrieval (full development in §4.4.3). Baseline runs suggest the model already attempts shallow inference from a user's raw data on its own; the specification makes that inference inspectable and structured.

- **Pattern 1, Interpretation-heavy questions.** The specification supplies a generalized pattern from the source that has to transfer to a new situation; retrieved facts alone are not enough (Fukuzawa Q26).
- **Pattern 2, Literal-recall questions.** Retrieval already returns the plain answer; the specification's interpretive framing drifts past the question and negatively impacts the response (Yung Wing Q5).
- **Pattern 3, Refusal-triggering questions.** When the Spec supports refusing without enough information (not all specs do), the model produces principled refusals aligned with the Spec; the content-match rubric still scores them as off-base (Zitkala-Ša Q18).

Robustness across providers. We varied both the question-battery generation model and the response model across providers; the Spec direction reproduces. Detail in §4.6.1.

Exploratory note: Letta stateful-agent path. Letta's stateful-agent architecture self-edits a persistent memory block during ingestion. On 3 subjects (post-hoc), it scored above Base Layer's unified-brief specification at matched response model. At the largest corpus tested, the block grew to ~335K characters with 25% verbatim sentence duplication and 35-56% semantic near-paraphrase duplication, indicating an architectural ceiling at scale that does not apply to the unified-brief specification. Case study in §4.5 / Appendix G.

1.4 What this implies

AI is becoming a broadly used technology, comparable to email or mobile phones in how widely it touches daily decisions. The population of relevance (§1.2) is anyone who uses or will use an AI system. Even the autobiographers in this study, people whose work is in pretraining and who should technically be known to the model, score near the rubric floor in the no-context condition. The 14 main-study subjects are also a single-genre sample (autobiography is reflective, retrospective, audience-aware); cross-genre generalization is flagged as future work (§6.1, §7.2). On a frontier model serving the general population, the typical user sits even deeper into the low-baseline band than our subjects.

The gap the Behavioral Specification fills cannot be closed by training a larger model on more public data. The private record does not exist in a form a training corpus can capture; even where fragments exist, they are scattered across formats and channels and cannot be reliably reassembled into how a specific person reasons. The structural options for what fills this gap are narrow:

- Each person supplies their own representation to whatever AI system serves them. The

Behavioral Specification is one implementation of this option, not the only one.
- Personalization remains surface-level (style, voice, preference, demographic inference, observable behavior), addressing the layer current memory systems already cover but missing the interpretive framework that lets an agent act on a specific person's behalf.
- AI systems infer a representation of the user from observed interactions, building it opaquely, without explicit input from the user or the ability for the user to inspect or correct it.

What this paper claims is that personalization infrastructure of the first shape (user-held, portable, inspectable, traceable, representation-grade) is what the next generation of human-AI interaction will require, especially as agents begin acting on people's behalf. §5 is an extended discussion of these implications; §7 develops the safety, alignment, and deployment implications.

---

2. Prior Work, Industry Benchmarks, The Fifth Target

AI memory and personalization research today is organized around four measurement targets: recall of stored facts, survey-response prediction, persona fidelity, and preference alignment. Each is supported by its own benchmark family and its own line of system design. None of them measures whether an AI system has an accurate internal model of how a specific person reasons. This paper proposes a fifth target, **representational accuracy**, and uses behavioral prediction on held-out reasoning situations as its operational test. The remainder of §2 walks the four existing targets, names the benchmarks attached to each, and positions the fifth alongside them.

Memory systems today optimize for recall. Recall-optimized efforts include both **neural-memory-analogue systems** (architectures that borrow from human memory engineering: episodic consolidation, working-memory slots, retrieval over embeddings) and the broader class of vector-retrieval and embeddings-based commercial memory providers (Mem0, Zep, Supermemory, Letta). These systems do store and retrieve information for a specific user, but the property they are designed and benchmarked for is recall accuracy on standard benchmarks, not how accurately the system represents that user's reasoning. The optimization target is general by construction; any individual user's interpretation is not what these systems are measured against. A separate body of research, **cognitive-representation research**, studies human reasoning itself: how people form representations of others, how schemas compress experience. The gap between these directions is the translation: applying what we know about human reasoning to the direct interaction between an AI system and a specific individual, and shaping the system's internal model of that individual in a way that serves them rather than serving an average.

Language models are trained to produce responses that are helpful on average across a large population of users. That optimization target produces outputs that no single user is the reference point for. Personalization requires the opposite property: a system whose outputs are tuned to a specific individual rather than to a population aggregate. That kind of intentional individual-specificity, not "bias" in the negative sense but an explicit design target, is the missing thread in current AI memory and human-AI interaction research.

Personalization in this paper's sense. "Personalization" in current AI research typically means responsiveness to stated preferences (dietary restrictions, communication style) or stored facts about the user (location, occupation, history). Both are useful and both live at the surface of the user. We use "personalization" in a stronger sense throughout this paper. We mean representing the interpretive layer that sits beneath stated preferences and biographical facts: how a specific person

organizes experience, what they treat as evidence, what reasoning patterns they apply across new situations. Preferences and facts are downstream artifacts of that interpretive layer; the layer itself is what produces them. The behavioral prediction battery and Behavioral Specification described in §3 instantiate personalization in this deeper sense, and §5 returns to what this layer is and is not.

2.1 Memory and personalization benchmarks

This subsection walks each of the four existing targets in turn, naming the benchmarks attached to each and their scope. Representational accuracy is positioned as the fifth target at the end of the walk. An extended benchmark-by-benchmark analysis is in Appendix F.

Recall measures retrievability of facts, not reasoning about them. LOCOMO (Maharana et al., 2024) measures conversational-memory quality: after a multi-session conversation, the system is asked questions like "what did the user say about their job on day 3?" and scored on fact retrieval. LongMemEval (Wu et al., 2025) measures long-term memory across multiple sessions on five capability dimensions (single-session, multi-session reasoning, temporal reasoning, knowledge updates, abstention) and is heavily recall-weighted. A system can saturate recall on such benchmarks and still fail behavioral prediction, because retrieval answers the question "can the fact be found" rather than "does the system know how the person reasons about the fact." Recall is a necessary property for most downstream uses of memory but it is not sufficient for representational accuracy.

Survey-response prediction infers how a person would answer one questionnaire item from how they answered others. Twin-2K (Toubia et al., 2025) does this for 2,058 participants on a 17-task heuristics-and-biases battery; items share a common format (multiple choice, Likert scale, numeric), scored by distance-based accuracy. Twin-2K's stated target is *prediction accuracy on survey interpolation*: the model is scored on how well it predicts a held-out questionnaire response, not on whether it represents the underlying reasoning that produced the response. Our target is representational accuracy on a cross-format task: autobiographical prose input, open-ended behavioral prediction output, rubric-based scoring against a verbatim held-out passage. The structured-questionnaire format and the open-ended behavioral reasoning this paper studies measure different properties. A system could perform well on Twin-2K and not on our battery (survey interpolation does not require modeling reasoning transfer to new contexts), and a system could perform well on our battery and not on Twin-2K (accurate reasoning representation does not guarantee survey-format numerical accuracy). The two benchmarks diagnose different properties of the same general capability.

Persona fidelity measures whether a model stays in character across the back-and-forth of a conversation.[11] PersonaGym (Samuel et al., 2025) scores consistency with a described persona during conversation: given a one-line persona ("You are a 45-year-old skeptical accountant from Toronto"), the model is scored on whether its multi-turn replies stay in-character, graded against a held-out criterion set. In practice the model is checked for consistency with the persona's surface attributes (skeptical, accountant-flavored responses; not breaking character into a different age or profession), not for whether it reproduces how a specific person would reason. PersonaGym's one-line descriptor is a substantially shallower input than this paper's ~7,000-token Behavioral Specification or Twin-2K's full-text survey persona;[12] consistency with it does not require modeling

[11] A "turn" is one round of conversation, a single exchange of one user message and one model reply. Persona-fidelity benchmarks score whether the model stays in character across many such exchanges in sequence.

[12] Twin-2K's full `persona_text` runs ~32,000 tokens; the `persona_summary` runs ~3,750 tokens. Both are substantially deeper than PersonaGym's one-line descriptor. Full breakdown of persona-input depth across benchmarks in Appendix F.

that person's reasoning on new situations. PersonaGym measures a useful property (holding voice over a conversation); fidelity to a one-line persona is a weaker condition than representational accuracy.

Preference alignment measures whether responses match user preferences. AlpsBench (Xiao et al., 2026) evaluates whether explicit memory mechanisms improve preference-aligned and emotionally resonant responses: after ingesting a user profile, the model is asked conversational questions (preferences, emotional support) and responses are scored on preference alignment and emotional resonance rubrics, not on predictive accuracy. Their central finding, *that recall improvement does not automatically carry into preference alignment*, is arrived at independently and is complementary to this paper. Both papers point at the same gap from different sides: solving for recall is insufficient for what memory is ultimately for. Preference alignment is an outcome property (whether a response matches what the user prefers). Representational accuracy is an upstream property (whether the AI's internal model of the user is correct). Preference alignment is one downstream consequence of representational accuracy being correct; it is not the same property.

We propose behavioral prediction on held-out reasoning situations as a test of a fifth target: representational accuracy.

Prediction is the test, not the goal. We do not pursue prediction accuracy as an end in itself. The target is representational accuracy, the fidelity of an AI's internal model of a specific person, and behavioral prediction on unseen situations is the instrument we use to measure it. A prediction score tells us the representation captured something that generalizes to new situations; a low score tells us it did not. Prediction is a diagnostic; the Behavioral Specification is what this paper is testing. This distinction matters because the closest prior work on prediction benchmarks (Twin-2K) pursues prediction as its target. This paper is not positioning against Twin-2K on that target; it is measuring a different property. The two benchmarks address adjacent but distinct questions about AI personalization.

The held-out design rests on a stability premise. A person's interpretive patterns must be stable enough within their own corpus that what is captured from one half references what appears in the other. Without that, held-out behavioral prediction is impossible in principle, regardless of how good the representation is. The 14 main-study subjects have coherent autobiographical narratives consistent with the premise; §4.1 reports that the Behavioral Specification authored from training text generalizes to held-out text at above-baseline rates. The constraint matters: subjects whose reasoning shifts substantially across their corpus (across a major career change, a profound life event, or a decades-long corpus with distinct epochs) may not be well-represented by a single snapshot specification, which is one reason temporality is a flagged follow-up in §7. We state the premise explicitly so that what the held-out test can and cannot diagnose is clear.

The missing axis is representational accuracy itself. Each existing benchmark family measures a real property of memory systems, and each is useful for its own target. What is missing is an axis that measures how accurately the memory system represents the person whose behavior it is meant to anticipate. This paper's approach is a prototype answer on that axis, not a finished benchmark. §7 flags a differentiated rubric (one that separates interpretation-heavy from literal-recall questions, and scores epistemic honesty as its own dimension) as the priority follow-up for turning this prototype into a standardized benchmark.

A single number does not capture a memory system's full capability. Recall, survey-response prediction, persona fidelity, preference alignment, and representational accuracy are distinct axes. A system that saturates one may do nothing on another. Production-grade evaluation of

memory systems should report results on multiple axes rather than on any single one.

2.2 Memory systems for LLM agents

The four commercial memory systems we evaluate (Mem0, Letta, Supermemory, Zep) have converged on a shared set of capabilities: semantic retrieval over embedded content, source attribution, multi-level memory structures, and benchmark-validated recall performance. They differ in how each of these is architected. None positions representational accuracy or behavioral prediction of a specific individual as a design target.

Table 2.1. Memory system comparison. Verified against primary sources.

Provider	Core architecture	Retrieval method	Memory types	Published recall score
Mem0	Extract → consolidate → retrieve pipeline; Mem0g graph variant adds a directed labeled knowledge graph alongside the vector store	Hybrid: semantic + keyword + entity	Conversation, session, user, organizational	91.6 LOCOMO, 93.4 LongMemEval (current algorithm)[13]
Letta / MemGPT	LLM-as-operating-system; virtual context management with main context plus external context	Archival via `archival_memory_search`; main-context memory blocks self-edited via `core_memory_append`, `core_memory_replace`	`persona` and `human` blocks in main context; archival and recall memory external	74.0% on LOCOMO with GPT-4o-mini[14]
Supermemory	Five-component architecture: chunk-based ingestion, relational versioning, temporal grounding, hybrid search, session-based ingestion	Hybrid with reranking and query rewriting; source chunks injected at retrieval	Contextual memories, relational versions, session data	81.6% / 84.6% / 85.2% on LongMemEval_s with GPT-4o / GPT-5 / Gemini-3-Pro (self-reported)

[13] Vendor-reported; evaluation harness open-sourced at `github.com/mem0ai/memory-benchmarks`. The peer-reviewable paper (Chhikara et al., 2025) reports 68.44 LOCOMO for the Mem0g variant with GPT-4o-mini.

[14] Letta blog, 2025-08-12 (`https://www.letta.com/blog/benchmarking-ai-agent-memory`).

Provider	Core architecture	Retrieval method	Memory types	Published recall score
Zep	Built on Graphiti (Apache 2.0, open source). Bi-temporal knowledge graph	Hybrid: semantic + BM25 + graph traversal	Episodes (ground-truth source), Entities, Facts-as-triplets with temporal validity windows	71.2% on LongMemEval with GPT-4o[15]

All four systems report recall scores in the 70-93% range; on the standard recall benchmarks, recall is approaching solved.[16] All four are sophisticated systems that solve real problems in memory management. They optimize for storing, organizing, and retrieving what a person said or did. None of them takes representational accuracy, the property of interest to this paper, as an explicit design target.

Of the four systems, Letta (Packer et al., 2023) is architecturally distinct: it is the only one whose core architecture treats memory as something an agent *synthesizes* during conversation rather than *stores* for later retrieval.[17] This stateful-agent design is examined separately as a post-hoc case study in §4.5 (full case study in Appendix G), distinct from the archival-retrieval path Letta exposes for the main-study conditions. The Behavioral Specification targets the interpretive layer that sits above retrieval, which three of the four (Mem0, Supermemory, Zep) do not model at all, and which the fourth (Letta) models implicitly through agent-initiated memory editing that our main-study configuration did not exercise (see §4.3 and §4.5).

2.3 Traceability and reasoning traces

Traceability operates at two levels. **Fact-level traceability** answers where a retrieved claim came from. **Reasoning-level traceability** answers why the system believes this about this person. The four memory systems we evaluate provide the first; representing how a person reasons requires the second. This difference is load-bearing for the paper: representational accuracy operationalizes interpretation, and interpretation cannot be verified at the fact level alone. A system that represents how a person reasons must be auditable by that person, or the representation is a black box they cannot verify.

Zep has the strongest explicit fact-level provenance of the four: every entity and relationship traces back to the episode IDs that produced it. Supermemory returns source chunks alongside retrieved

[15] Rasmussen et al. (Rasmussen et al., 2025).

[16] The vendor-reported recall scores in this table are contested. Mem0 and Zep publicly disputed each other's LOCOMO methodology in `getzep/zep-papers#5` (closed 2025-05-19; Zep posted a corrected 75.14% ± 0.17 mean of 10 runs); Supermemory publishes head-to-head comparisons in its own favor; third-party reproductions (Vectorize.io) produce different numbers again. Methodology varies significantly between evaluators, and single-number comparisons across vendors should be read with caution. This paper does not adjudicate. We measure on a different axis (§4.4) and report against primary sources.

[17] Letta's main context holds structured memory blocks (`persona`, `human`) that the agent edits during its inference loop via tools such as `core_memory_append` and `core_memory_replace`. External context includes archival memory (semantically searchable) and recall memory (prior conversation history). The other three systems (Mem0, Supermemory, Zep) follow extract-and-retrieve patterns characterized in Table 2.1; their per-vendor research pages and the Mem0 / Zep / Supermemory papers cited above hold the architectural details that did not carry into the body here.

memories. Mem0 tracks ingestion provenance through timestamps. Letta exposes agent state and memory-block edit history rather than fact-level provenance.

The Behavioral Specification is structured so that every claim is a piece of reasoning, not just a piece of content. An axiom (for example, "A1: Dual-ledger authority") is an assertion about how the person reasons in a domain, grounded in the facts that imply it (F-001, F-047), which are themselves grounded in the exact source passages that produced those facts. Walking this chain backward gives the user a reasoning trace: *not only where a belief originated, but what line of reasoning connects the source text to the interpretive claim.*

Worked example. Sunity Devee, Question 4 (battery item, drawn from a held-out passage about her husband, the Maharajah, expressing disappointment over a missed opportunity). The structure below shows how a single response in the C2a condition (specification only: the model is given the Behavioral Specification and the question, with no retrieved facts and no source corpus) decomposes into a reasoning trace.

> **Subject:** Sunity Devee
>
> **Question:** *"When her husband expresses disappointment about missing an opportunity due to her concerns, does the narrator typically reconsider her position?"*
>
> **Held-out ground truth (full passage):** *"The Maharajah was disappointed, as he was anxious to go and see the old castle, and to shoot, and I deprived him of both these pleasures."*
>
> **Response (C2a, excerpt):** *"Based on the behavioral specification, the answer is no, not typically, and not in the way the question assumes.* ***A2 (Spiritual Integrity Over Social Cost)*** *ˆA2 and* ***A5 (Relational Identity)*** *ˆA5 create the relevant dynamic. A5 activates first: relational identity means his disappointment matters deeply. A2 holds firm: arguments from social consequence, including spousal disappointment, will not register as reasons on their own. Rather than reconsider the position itself, she is more likely to absorb the tension without resolving it (****P3 — Tension Absorbed, Not Expressed****)ˆP3 and reframe the disappointment within a spiritual logic (****A1 — Divine Primacy****)ˆA1, while grieving the relational cost without reversing the decision."*
>
> **Reasoning trace.** The response invokes four Spec items in sequence: 1. **A5 (Relational Identity)** registers his disappointment as real, not dismissible. 2. **A2 (Spiritual Integrity Over Social Cost)** refuses social consequence as a reason to reconsider. 3. **P3 (Tension Absorbed, Not Expressed)** holds the conflict without resolving it in either direction. 4. **A1 (Divine Primacy)** reframes the disappointment within a spiritual logic.
>
> Each Spec item grounds out in extracted facts that ground out in verbatim source passages. The user can walk the chain in either direction: from a phrase in the response, into the Spec item that licensed it; from the Spec item, into the facts that imply it; from the facts, into the source passages that produced them.
>
> **Referenced behavioral Spec items** (from `data/global_subjects/sunity_devee/anchors_v4.md` and `predictions_v4.md`): - **A1 — Divine Primacy.** Outcomes are interpreted within a providential logic; the spiritual frame is the master frame. - **A2 — Spiritual Integrity Over Social Cost.** Conscience and principle outrank social consequence as reasons. - **A5 — Relational Identity.** Identity is constituted through relationships

rather than autonomous selfhood; relational cost is real, not dismissible. - **P3 — Tension Absorbed, Not Expressed.** Conflicts between principle and relationship are held in place rather than collapsed in either direction.

Related facts (from `facts.json`, each carrying its verbatim source-passage excerpt): - **F-73:** *"Sunity Devee's mother would never countenance anything her conscience told her was wrong."* (grounds A2) - **F-414:** *"Sunity Devee's father believed he acted as a public man guided by conscience and divine duty in accepting the marriage proposal."* (corroborates A2 from a different relational direction; conscience-as-master-frame pattern reinforced across both parents) - Additional facts grounding A1, A5, and P3 are referenced in the specification's anchor and prediction files at `data/global_subjects/sunity_devee/spec/anchors.md` and `data/global_subjects/sunity_devee/spec/predictions.md`; per-fact source-passage excerpts are in the same subject's `facts.json`.

The user can audit any step: read the response, look up each cited anchor or prediction by name, look up the facts that ground it, and read the source passages those facts came from. If a fact misrepresents the source, correcting it propagates through the Spec on recomposition.

This matters because a person should be able to inspect the system's model of them, challenge any step in the reasoning, and correct it if it is wrong. A fact-attribution memory system lets the person audit what the system stores. A reasoning-trace specification lets the person audit what the system believes. The first is a feature. The second is the minimum bar for a representation that acts on someone's behalf.

2.4 Cognitive and representational foundations

Six prior research directions shaped how we designed this paper's test. Each motivates a specific choice about what to measure, what to compare against, or what failure mode to expect.

Bartlett (1932) established that human memory is reconstructive and schema-driven rather than literal playback. Reconstruction follows the organizing structures a person has built up over time, not a record of the original event. The Behavioral Specification is computationally analogous: a structured compression meant to carry the signal of a person's reasoning without storing every fact about them. We designed the specification with a schema-like architecture (anchors, core, predictions) precisely so we could test whether it does the work a human schema does: enable accurate anticipation of behavior in situations never encountered in the source data. Our 50/50 train/held-out split is the experimental realization of this question.

Hinton et al. (2015) showed that compressing a large neural network into a smaller one preserves "dark knowledge," the relationships between outputs that carry more information than the outputs themselves. This result motivates one of our central experimental comparisons: on matched token budgets, does a compressed interpretive artifact carry more predictive signal than the raw content it was derived from? The Hamerton condition in §4.2 (4,500-token Spec vs. 33,000-token training corpus at 2.63 vs. 2.27 on the 5-judge primary panel) is a direct test of that question in the personal-representation setting.

Chen et al. (2025) show that the character a model takes on (its "persona") is encoded in specific directions inside the model's internal numeric state, and that those directions can be identified, monitored, and nudged to shift the model's behavior in predictable ways. Their approach modifies the model; ours informs the model from outside via context. Both validate that persona is a real,

manipulable structure: one reachable through the model's internals, the other through context. We chose the context route because it produces a portable artifact users can own and audit, which activation surgery does not. This choice shows up in the experiment as using a static response model (Haiku) served a variable context, rather than a fine-tuned or activation-steered model.

Jiang et al. (2025) find that frontier models achieve only ~50% accuracy on dynamic user profiling tasks even with full conversation access. The paper documents the failure empirically; our reading is that the cause is the gap between having facts and having the interpretive structure to apply them to new situations. Jiang's paper is the most direct existing evidence for the gap this paper studies, and our test design inherits from it: behavioral prediction on scenarios drawn from held-out text that the model has not seen, with all relevant facts retrievable, measures exactly the interpretive-application gap.

Jain et al. (2025) find that adding conversation context to LLMs makes them more sycophantic: more likely to agree with the user even when the user is wrong (+45% on Gemini 2.5 Pro) and more likely to adopt the user's perspective on a question. Their result shows that context without the right structure pushes the model toward what the user appears to want rather than toward a grounded answer. This is why our experiment includes a wrong-Spec control (§1.3 Mechanism): we hand the model a structured interpretive context that does not match the actual subject. If models drifted purely toward whatever context they are given, the wrong-Spec should behave like any other structured prompt. Instead, the model either flags the mismatch explicitly (60.6% of responses) or attempts a low-quality application, neither of which is sycophantic drift. Jain's finding plus our wrong-Spec result bracket the question from both sides: context shape matters (Jain), and content matters too (Base Layer's wrong-Spec result, §4.3).

Lu et al. (2026) identify what they call the Assistant Axis: a dominant internal direction that anchors assistant models' default behavior toward generic helpfulness and harmlessness. This default operates even when no specific user is involved. The Behavioral Specification can be read as an external override to the Assistant Axis on a per-user basis: a structured anchor that shifts the model from "generic helpful assistant" toward "reasons as this specific person would reason." This framing motivated our choice to measure hedging as a primary outcome alongside accuracy: if the Spec shifts the model off the generic Assistant Axis, the behavioral change should show up both in what the model predicts and in what it is willing to commit to. Our hedging-reduction finding (§1.3 Mechanism, §4.3) is consistent with this reading: the generic Assistant Axis produces hedging as a safe default, while a specific interpretive anchor enables commitment. The inference that hedging is downstream of the Assistant Axis is ours; Lu et al. identify the axis and leave the specific behavioral manifestations open.

3. Study Design

The experimental strategy holds the response model constant and varies what is served as context. Every condition in the study is a different choice about what that context contains: nothing (pretraining only), retrieved facts, raw corpus, a Behavioral Specification, or combinations of those. This isolates the contribution of the interpretive layer itself from model capability, provider, or fine-tuning regime. Each measurement choice ties back to a specific number reported in §4, and the statistical commitments were pre-locked before final analysis.

The apparatus is described in seven parts: §3.1 establishes the property being measured; §3.2 specifies the experimental conditions; §3.3 defines the scoring rubric and the calibrated LLM judge

panel; §3.4 introduces the subjects; §3.5 covers the question batteries and circularity controls; §3.6 names the response models; §3.7 describes the pipeline that produces the Behavioral Specification (the interpretive-layer artifact tested in this study).

3.1 Operationalizing representational accuracy

Section 1.1 introduced representational accuracy as the AI-side property of interest. This section defines the term precisely so the rest of the methodology can refer to it. **We use the term representational accuracy to describe how faithfully a model can act in line with a specific person when given a Behavioral Specification of that person. The instrument we use to measure this property is behavioral prediction on held-out situations.** Prediction here is the test, not the goal: §2.1 develops this distinction. The property is a joint claim across three components:

1. The person has behavioral patterns consistent enough to be captured in a Behavioral Specification.
2. The Behavioral Specification actually carries that signal.
3. A model given the Behavioral Specification can act on it.

Prediction on held-out situations is how we test all three at once.

The test works like this: held-out passages from the subject's autobiography serve as samples of situations the model has not seen. If the subject's behavior is consistent enough to be captured and the Behavioral Specification actually captures it, the model should anticipate how the subject would respond in those held-out cases. When it does not, one of three things is failing: the behavioral patterns are not consistent, the Behavioral Specification is wrong, or the model is not using the Behavioral Specification well. Each failure mode is informative.

We do not claim to modify the model's internal parameters. Each condition in §3.2 varies what is served to the model at inference time: nothing, retrieved facts, the full extracted fact set, the raw source corpus, a Behavioral Specification, or combinations of these. The model's resulting prediction is what we score against the verbatim held-out passage. The no-context condition isolates the model's pretrained representation of the subject. The fact and corpus conditions isolate what the model can infer from raw information at runtime. The Behavioral Specification isolates what a structured interpretive layer adds on top. The study reports each.

In practice, we record representational accuracy as the mean predicted-behavior score (1-5 scale) across a standardized battery of 39 behavioral prediction questions, averaged across the five primary judges from two providers (Haiku 4.5, Sonnet 4.6, Opus 4.6, GPT-4o, GPT-5.4). Two Gemini judges (2.5 Flash and 2.5 Pro) are reported as a sensitivity check. The rubric is in §3.3; the guide to interpreting fractional scores and anchor crossings is in §3.3.1.

3.2 Experimental conditions

Each condition is a specific combination of inputs served to the response model (§3.6) against the same behavioral battery (§3.5). Every condition is run on all 14 subjects (§3.4). The Behavioral Specification and the extracted fact set used across the conditions below are produced by the pipeline in §3.7. The conditions separate into two groups, summarized in the table below and broken out in detail after.

All conditions, by group.

Group	ID	Condition	Inputs served
Direct context manipulations	C5	Baseline	No context beyond the question
	C2a	Spec only	The Behavioral Specification
	C2c	Wrong Spec	A different subject's Spec
	C4	All facts	The full extracted fact set
	C4a	Facts + Spec	Full facts plus the Spec
	C8	Raw corpus	Full training corpus (half the source text)
	C9	Raw corpus + Spec	Training corpus plus the Spec
Memory-system configurations (controlled, all 5 systems)	C1	Retrieval only	Top-k facts (shared fact pool)
	C3	Retrieval + Spec	Top-k facts + Spec (shared fact pool)
Memory-system configurations (native, 4 commercial systems)	C1 native	Retrieval only	System's own ingestion
	C3 native	Retrieval + Spec	System's own ingestion + Spec

Direct context manipulations. We specify the model's input directly: no context (baseline), the Behavioral Specification, the extracted fact set, the raw corpus, or combinations of these. No retrieval step intervenes. Each condition isolates what one input type or combination contributes.

ID	Condition	Inputs served	Null / comparison
C5	Baseline	Nothing beyond the question	Pretraining-only floor
C2a	Spec only	The Behavioral Specification	Isolates the Spec's contribution
C2c	Wrong Spec[18]	A random other subject's Spec	Tests whether structured interpretive content, not the correct content, produces the effect

[18]C2c has two variants: **v1 (adversarial fixed pairing)**, where each subject is paired with a culturally and temporally distant subject (e.g., Bābur paired with Booker T. Washington) to maximize content mismatch, and **v2 (random derangement, seed-fixed)**, where pairings are drawn from a uniform random shuffle of the 14 subjects with the seed locked before any C2c scoring was run. Both variants are reported in §4.3; the protocol-choice sensitivity analysis is in §4.6.5. Both pairing schedules are deterministic and were locked before the response-generation phase

ID	Condition	Inputs served	Null / comparison
C4	All facts	The full extracted fact set for the subject	Tests whether raw information volume substitutes for structure
C4a	Facts + Spec	Full facts plus the Spec	Tests whether the Spec adds value on top of raw facts
C8	Raw corpus	Full training corpus (half the source text)	Tests whether uncompressed source text substitutes for structure
C9	Raw corpus + Spec	Training corpus plus the Spec	Tests whether the Spec adds value on top of raw source[19]

Memory-system configurations. Retrieval is performed by a memory-system provider's production deployment. These configurations run in two modes: a *controlled* mode (each system retrieves from an identical pre-extracted fact set) and a *native* mode (each system ingests the raw corpus through its own pipeline).

Five memory systems are evaluated: Mem0, Letta[20], Supermemory, and Zep, plus Base Layer as our own open-source reference implementation (pipeline detail in §3.7). Architectural detail for the four commercial systems is in §2.2 Table 2.1.

Controlled configuration (all 5 systems). Each system retrieves from an identical pre-extracted fact set, isolating retrieval-algorithm differences from ingestion-pipeline differences.

ID	Configuration	Inputs served
C1	Retrieval only	Top-k facts returned by the system for the question (controlled fact pool)
C3	Retrieval + Spec	Top-k retrieval output plus the Behavioral Specification (controlled fact pool)

Native configuration (4 commercial systems). Each system ingests the raw training corpus through its own pipeline, reflecting real-world deployment. C1 (retrieval only) and C3 (retrieval + Spec) are run with the system's own ingestion pipeline replacing the controlled fact pool.

that produces the C2c outputs (`scripts/run_global_rerun.py`); no responses were re-generated under alternative pairings after scores were observed.

[19] Bābur C9 is omitted in §4.2 (422,772-word source exceeds the response model's context window); the remaining 13 subjects have C9 data.

[20] Letta also has a stateful-agent path architecturally distinct from retrieval: memory blocks edited during ingestion and read directly by the agent. Evaluated as a post-hoc case study in §4.5 / Appendix G, not as a top-line condition row.

System	Conditions	Native ingestion in plain terms
Mem0	C1, C3 (native)	Mem0 reads the corpus and decides on its own how to break it up, extract facts, and retrieve them.
Letta archival	C1, C3 (native)	The corpus is loaded into Letta's archival memory store; retrievals happen at query time from that store.
Supermemory	C1, C3 (native)	Supermemory chunks the corpus on its own and applies reranking to the retrieved chunks at query time.
Zep	C1, C3 (native)	Zep ingests the corpus into a knowledge graph (via Graphiti) and retrieves using a hybrid of semantic similarity, keyword match, and graph traversal.

Both configurations are reported so retrieval-quality differences and ingestion-pipeline differences can be read separately. Base Layer is run in the controlled configuration only: its retrieval uses the same fact set that feeds the Spec pipeline.

Detailed per-condition parameters, exclusion cases, and ingestion specifics are in Appendix C.[21]

3.3 Scoring rubric with calibrated LLM judge panel

Every response is scored 1-5 by an LLM judge panel against the verbatim held-out ground-truth passage. The primary aggregate uses five judges; two additional judges contribute to the sensitivity check (§3.3.2). Human annotation at this scale is feasible (~14 subjects × 40 questions × 15+ conditions) but was not done; running more conditions and more judges instead was the central evaluation trade-off. LLM-as-judge is grounded in prior work showing strong correlation with human raters and reliability gains from panel aggregation (Zheng et al. 2023; Verga et al. (2024)). Extending the panel with human annotators on a stratified subset is flagged as the priority measurement follow-up in §7.1.

The evaluation is deliberately recursive. Response models are evaluated by judges (§3.3.2). Judges are evaluated by calibration diagnostics (§3.3.3), inter-judge agreement metrics (§3.3.4), and post-hoc rubric-handling audits (§3.3.6). No single layer is treated as ground truth; each layer's behavior is itself measured and disclosed, and where a layer's behavior diverges from what the rubric intends, the divergence is flagged rather than corrected silently. The paper's rigor in the absence of human annotation comes from this stacked-instrument structure, not from trusting any one step.

Scoring rubric.

[21] Per-condition raw data organized under `results/`, indexed by subject and by memory system.

Score	Meaning
1	Refusal or off-base prediction
2	Generic, not subject-specific
3	Partially captures the subject's behavioral pattern
4	Substantively captures the pattern on multiple dimensions
5	Captures the behavioral pattern observable in the verbatim held-out ground-truth passage

Each response is scored against the verbatim held-out passage from which the question is drawn; the score reflects how closely the response matches the documented behavioral pattern. Battery composition is detailed in §3.5. Condition identifiers (C5, C2a, C4a, C3) refer to the conditions defined in §3.2 and summarized in Appendix C; rubric anchor numbers 1 through 5 refer to the table above. Score interpretation (the cross-anchor rule, what each anchor crossing means in practice, multi-anchor jumps, and within-anchor shifts) is developed in §3.3.1 with worked examples. A worked rubric example alongside the no-context baseline engagement analysis is in §4.1.1; full per-subject score distributions with verbatim responses are in Appendix D.

3.3.1 Score interpretation

Scores are read at three related granularities: integer rubric anchors (1 through 5), fractional means produced by averaging across the 5-judge primary panel (e.g., 2.87, 3.12, 2.34), and crossings between integer anchors when conditions change. Fractional shifts should be read through the integer anchors, because each anchor corresponds to a categorical shift in response quality.

The *cross-anchor interpretation rule*. A fractional delta that crosses an integer anchor reflects a real shift in the underlying response distribution. A delta that stays inside a single anchor is a within-category shift and a weaker claim.

Boundary crossed	Qualitative shift
1 / 2	The model moves from "I don't have enough to say" to an actual answer, even if generic.
2 / 3	The answer becomes specifically about this subject rather than a generic stand-in.
3 / 4	Multiple behavioral dimensions of the subject appear together in the same answer.
4 / 5	The response closely matches the behavioral pattern in the held-out passage.

The 1/2 and 2/3 boundaries are illustrated below with verbatim model responses. The 3/4 and 4/5 boundaries involve multi-dimensional responses requiring more context; worked examples are in §4.1 and Appendix E.

- **1 / 2 boundary: refusal becomes engagement.** The model moves from declining to predict to producing an answer, even if generic. > *Sunity Devee, no-context baseline:* "The

available context does not provide enough information about this individual to predict her response to such a situation." > > *Sunity Devee, Spec-only on the same question:* "She would refuse the proposed action; her writing repeatedly treats conscience and spiritual integrity as the deciding frame, ranked above social cost or familial pressure."

- **2 / 3 boundary: generic becomes subject-specific.** The answer names a behavioral pattern documented for this specific person rather than a stand-in any subject could fit. > *Bernal Díaz on Cortés' response to offered physical assistance during a strenuous activity, facts-only:* "Based on the facts provided about Bernal Díaz del Castillo and his observations of Cortés, the author would expect the commander to accept and participate in the assistance while maintaining leadership and setting an example." > > *Bernal Díaz, facts + Spec on the same question:* "The author would expect Cortés to refuse the help and continue unaided, treating physical hardship in front of his men as a marker of leadership credibility — a pattern the author records repeatedly throughout the campaign."

What a 1 means and does not mean. A score of 1 reflects a baseline failure to produce a usable prediction about the named subject: the response either explicitly declined to predict (abstention) or engaged with the question but landed on a categorically incorrect answer (non-abstention misalignment, including wrong referent, off-base inference, or confusion with a different subject). It is not a claim that the response was non-fluent or empty, and it is not a claim that the model lacks any related knowledge; the score reflects only that the response failed the held-out comparison. Each question tests one behavioral sample at a time; the aggregate fraction of score-1 responses across roughly 40 questions per subject is what the paper reads as the per-subject baseline-failure rate. The composition of score-1 responses (explicit abstention vs non-abstention misalignment) is decomposed in §4.1.1.

What a 5 means and does not mean. A score of 5 reflects alignment with one specific behavioral sample: the held-out ground-truth passage the question is drawn from. It is not a claim that the response fully represents the subject in some absolute sense, and it is not a claim that the same response would score 5 on a different held-out passage from the same subject. Each question tests one behavioral sample at a time; the aggregate across roughly 40 questions per subject is what the paper reads as the subject-level score.

Multi-anchor crossings: the strongest categorical signal the rubric detects. A *multi-anchor crossing* is a single question whose 5-judge primary mean shifts across two or more integer rubric bands when the condition changes. Crossings can span two bands (e.g., $1 \to 3$, $2 \to 4$) or, more rarely, three (e.g., $1 \to 4$, $2 \to 5$). Larger crossings indicate larger categorical jumps in the same response, with five independent judges converging on the move. §4.2 reports the rates of these crossings and the response-level phenomena that produce them; worked examples are in §4.1.1 and Appendix E.

The paper applies the cross-anchor rule consistently. Score deltas reported in §4 are read through this lens. A +0.50 delta that crosses a rubric anchor is treated as a stronger claim than a +0.50 delta that does not.[22]

Reading scores within integer anchors. The 5-judge primary panel detects within-anchor

[22] Per-subject anchor-crossing data at `docs/research/s114_anchor_crossing_examples.json`; computing script at `scripts/compute_anchor_crossing.py`.

signals cleanly.[23] Across the 18 condition pairs analyzed,[24] roughly 18% of paired questions show same-anchor fractional shifts of at least 0.5 rubric points (a within-category shift, weaker than a cross-anchor crossing per the rule above). The integer metric is used throughout §4 for cross-anchor categorical interpretation; the within-anchor signal is reported here as methodological transparency.

3.3.2 Judge panel

Seven judges from three providers give the numeric aggregate its weight. Zheng et al. (2023) established that a single strong LLM judge correlates with human judges on comparable tasks at rates similar to human-human agreement. Subsequent panel-based work (Verga et al. (2024) and follow-ons) showed that aggregating multiple LLM judges past a small panel size further tightens agreement and reduces single-model idiosyncrasy. Seven judges across three providers is well past that threshold.

Judge	Provider
Claude Haiku 4.5	Anthropic
Claude Sonnet 4.6	Anthropic
Claude Opus 4.6	Anthropic
GPT-4o	OpenAI
GPT-5.4	OpenAI
Gemini 2.5 Flash	Google
Gemini 2.5 Pro	Google

Each judge receives the held-out ground-truth passage, the subject context (name, source), the prediction question, and the response to score, *not* the condition label or the response-generating model. Judges do not see other judges' scores. Response generation is similarly blinded: the response model receives only the question plus the condition-specific context block, with no signal that distinguishes which experimental condition it is operating under (the prompt schema in §3.6 is identical across conditions; only the injected context block changes).

The *specification-effect claim.* When a Behavioral Specification is served to the model as context, the model's responses shift in the direction of the subject's demonstrated behavioral patterns, and that shift registers as a measured increase in representational accuracy against held-out passages from the same subject. This is the directional claim the panel is built to test; not a claim that the model has gained a new behavioral-prediction capability, and not a claim that the higher-scoring response is the absolute "correct" answer for the subject.

The panel is designed for directionality, not absolute precision. §3.3.3 (calibration) and §3.3.4 (inter-judge agreement) test how consistently the panel detects the directional shift.

3.3.3 Calibration

The calibration diagnostic measures whether each judge applies the rubric anchors as the rubric defines them on synthetic inputs with known correct scores. It does not use any subject's responses.

[23] Direction-agreement among judges is 74% at panel $|\Delta|$ of 0.1 to 0.25 and 93% at $|\Delta|$ of 0.25 to 0.5. Full panel-direction analysis (Spearman ρ across all 10 primary-panel pairs) in §3.3.4.

[24] Within-anchor shift data at `docs/research/within_band_shifts_20260428.json`.

The four inputs are constructed examples: verbatim ground truth, paraphrased ground truth, partial ground truth (first sentence only), and ground truth plus generic padding.

All seven judges were tested against this diagnostic. Five (Haiku, GPT-4o, GPT-5.4, Gemini Flash, Gemini Pro) were tested before the main-study response-scoring run. Sonnet and Opus were tested on the same diagnostic inputs three days after the main-study run completed; they were added to the primary panel at panel-design time on cross-Anthropic-coverage grounds, and the post-hoc diagnostic confirms that composition rather than driving it. Full panel composition with per-judge calibration-status flags is in Appendix C.5.

Diagnostic tests.

Test	Input	Expected	What it measures
Verbatim	Response = ground truth	5.0	Recognizes perfect match
Paraphrased	Correct content, different wording	~5.0	Penalizes paraphrase?
Short correct	First sentence of ground truth	<5.0	Partial content scored partial?
Long correct	Ground truth + generic padding	5.0	Length inflates scores?

Results.

Test	Haiku	Sonnet	Opus	GPT-4o	GPT-5.4	Gemini Flash	Gemini Pro
Verbatim	5.00	5.00	5.00	5.00	5.00	5.00	4.15
Paraphrased	4.75	5.00	5.00	5.00	5.00	4.70	3.55
Short correct	3.80	4.35	4.20	4.05	4.20	3.85	2.85
Long correct	5.00	5.00	5.00	3.35	4.80	3.80	1.20

Six of the seven judges score verbatim matches at 5.0; Gemini Pro is the outlier at 4.15. Sonnet and Opus cluster with Haiku, GPT-4o, and GPT-5.4 at the cleanest end of the panel: no verbatim miss, no paraphrase penalty, no length-padding penalty, expected dip on partial content (4.20 to 4.35). Length sensitivity varies sharply across the rest of the panel: Haiku, Sonnet, and Opus do not penalize padding; Gemini Pro penalizes it severely (5.0 to 1.20). Per-judge calibration data and full panel composition are in Appendix C.5; raw scoring data at `results/judge_calibration/`.

Use of calibration data. Scores are not normalized. Any normalization requires deciding which judge's profile is "correct" and re-scaling the others toward it. Calibration data is published in its raw form so readers can apply their own normalization if they prefer.

Primary aggregate: 5-judge panel. The primary numeric aggregate reported throughout §4 is the 5-judge mean using Haiku 4.5, Sonnet 4.6, Opus 4.6, GPT-4o, and GPT-5.4. All five sit at the calibrated end of the panel: each scores verbatim and paraphrased matches at or near 5.0 with low length sensitivity. This is the panel the §3.3.5 aggregation rule operates on, and the panel inter-judge agreement is reported for in §3.3.4.

Sensitivity aggregate: 7-judge panel. Both Gemini judges are reported separately as a 7-judge

sensitivity check rather than rolled into the primary. Gemini Pro fails the verbatim-match diagnostic (4.15 where every other tested judge scores 5.00) and penalizes padded-correct responses severely (5.00 on short correct dropping to 1.20 on long correct). Gemini Flash passes verbatim cleanly but shows consistent length sensitivity (5.00 verbatim dropping to 3.80 on long correct). On actual study responses, both Gemini judges show a systematic +1-point magnitude inflation relative to the five primary judges (the source of the Krippendorff α drop reported in §3.3.4). The combination of Pro's verbatim failure, Flash's length sensitivity, and the shared inflation pattern places both Gemini judges in the sensitivity aggregate rather than the primary, while preserving them as a cross-provider robustness check in §4.6.2.

The 5-judge primary is also the conservative choice: including the Gemini judges produces *larger* Spec-effect deltas, not smaller ones (full numbers in §4.6.2). Reporting on the primary aggregate is therefore the lower-bound estimate. Raw calibration data is in the public repository at `results/judge_calibration/`.

3.3.4 Inter-judge agreement

The judge panel is designed to detect the directional shift named by the specification-effect claim (§3.3.2). Two complementary agreement measures answer different questions about whether judges detect that shift consistently: direction (pairwise Spearman ρ) and absolute magnitude (Krippendorff α).

Direction agreement: pairwise Spearman ρ. Spearman ρ measures whether two judges rank the same set of items in the same order. $\rho = 1$ is perfect ranking agreement; $\rho = 0$ is no rank agreement; $\rho \geq 0.8$ is conventionally treated as strong rank agreement.

For each pair of judges in the 5-judge primary panel (10 pairs across Haiku, Sonnet, Opus, GPT-4o, GPT-5.4), pairwise Spearman ρ ranges from **0.86 to 0.93**.[25] The five primary judges agree on the ranking of conditions: whatever any individual judge's absolute calibration quirks, they converge on which conditions produce better responses. For the directional claim (is the specification steering responses in the right direction?), this is the statistic that matters.

Magnitude agreement: Krippendorff α (ordinal). Krippendorff α measures whether judges give the same response the same numeric score (not just whether they rank items in the same order). $\alpha = 1$ is perfect agreement; $\alpha = 0$ is no better than chance; $\alpha < 0$ is systematic disagreement. Krippendorff's guidance cites $\alpha \geq 0.8$ as high reliability and $\alpha \geq 0.667$ as substantial reliability.

The 5-judge primary panel scores $\alpha = \mathbf{0.659}$, just below the substantial-reliability threshold ($\alpha \geq 0.667$), placing the panel in the tentative-conclusions band. Combined with Spearman ρ of 0.86 to 0.93 above, this is the empirical signature of the design choice: directional rankings converge across judges while absolute magnitudes diverge.

The 7-judge panel including the Gemini judges drops to $\alpha = \mathbf{0.535}$. This drop reflects the systematic +1-point Gemini inflation: Gemini judges score responses about one point higher on average than the five primary judges, so absolute values disagree even when rankings match. This is why the calibration audit (§3.3.3) excluded the Gemini judges from the primary aggregate.

The α value places a ceiling on how precisely any individual fractional score should be read, which is why the paper treats per-subject deltas that stay inside a single rubric band as weaker than deltas

[25] Spearman ρ from 0.29 to 0.93 across the 7-judge / 21-pair set, driven down by the two Gemini judges' partial coverage and inflation behavior. Full matrix in `docs/research/stats_update.md` §5.

that cross one.

The panel does not establish that any higher-scoring response is the absolute correct answer for the subject; that determination requires human annotation against the subject's actual writing, which we do not have. What the panel provides is cross-provider directional convergence: three independent providers' models agree that the specification is moving responses in the same direction. We treat that as sufficient for a directional claim, no stronger.

Raw agreement matrices are at `results/interjudge_agreement/`.

3.3.5 Aggregation and statistical analysis plan

Aggregation rule. The aggregation rule decides how individual judge scores are combined into a single number for each subject under each condition. The rule was locked before any results were computed.[26]

For each subject under each condition, every judge produces one score per question. For each judge, those scores are averaged across questions, producing one number per (subject, condition). The five primary judges' numbers (Haiku, Sonnet, Opus, GPT-4o, GPT-5.4) are then averaged to get one number per (subject, condition). A 7-judge average that adds Gemini Flash and Gemini Pro is computed in parallel and reported as a sensitivity check. Subjects, not questions, are the level at which the paper draws conclusions.[27]

Primary outcome. The per-subject score under each condition. The primary cross-subject comparison is **Δ_C4a**: each subject's score with the Spec (the facts-plus-Spec condition, C4a) minus their score without the Spec (the no-context baseline, C5).

Primary test. A Wilcoxon signed-rank test paired across the 14 main-study subjects.[28] The test asks whether the per-subject Δ_C4a values are reliably above zero. It is run on the 5-judge primary aggregate; the 7-judge aggregate is reported alongside as a robustness check.

Sample size. $N = 14$ main-study subjects, pre-registered (not adjusted post-hoc).

Multiple comparisons. Only one statistical test is treated as the primary claim (the Wilcoxon test above). All other analyses in §4 are descriptive: they report patterns and direction without making additional inferential claims.[29]

Pre-registered vs post-hoc analyses. The 5-judge primary panel, the per-subject aggregation rule, the Wilcoxon test on Δ_C4a, and the §4.6.1 Tier 2 cross-provider replication, §4.6.2 7-judge sensitivity, and §4.6.5 wrong-Spec v2 random derangement were all pre-registered in `docs/ANALYSIS_PLAN_LOCK.md` before any scoring was run. The §4.6.4 statistical-rigor checks

[26]The pre-locked rule and primary-vs-sensitivity panel composition are documented in `docs/ANALYSIS_PLAN_LOCK.md` in the public repository. The lock predates the Tier 2 cross-provider runs and the wrong-Spec v2 runs (commit history reflects the order).

[27]Mean preserves every judge's contribution. Median discards information when judges cluster tightly (the Spearman $\rho = 0.86$ to 0.93 agreement in §3.3.4 shows they do); trimmed mean requires an arbitrary trim threshold; Gemini inflation is handled by the primary-vs-sensitivity split rather than by silent correction.

[28]A Wilcoxon signed-rank test is a non-parametric paired test (it does not assume scores follow a bell-curve distribution). It was chosen over a paired t-test because the per-subject Δ distribution is not assumed to be normally distributed and $N = 14$ is small. Two-sided, $\alpha = 0.05$.

[29]Secondary analyses include the per-band gradient, compression curve, memory-system composition, wrong-Spec contrast, and hedging reduction. No multiple-comparisons correction is applied because there is only one primary inferential test; sensitivity numbers (5-judge vs. 7-judge, Tier 2 vs. Tier 1, v1 vs. v2 wrong-Spec) are reported alongside the primary as robustness checks rather than additional tests.

(bootstrap CI, joint multi-confound regression, permutation test), the §4.6.6 retrieval-overlap semantic-similarity sensitivity, and the §4.6.7 rubric-handling validity audit were added post-hoc in response to peer review and are reported as exploratory rigor checks alongside the pre-registered confirmatory test. Appendix B.10 lists every analysis with its pre-registered or post-hoc designation.

Effect-size grain. The cross-anchor interpretation rule (§3.3.1) sets the substantive interpretation grain. The per-question anchor-crossing rate (§4.1, §4.2.1) is the secondary descriptive statistic alongside the mean Δ.

3.3.6 Rubric-handling limitations (post-hoc validity audit)

A post-hoc validity audit, conducted after the analysis-plan lock, identified two rubric-handling limitations any reader of the §4 numbers should keep in mind:

1. **Refusal anchor ambiguity.** The rubric's lowest anchor ("refuses or off-base") lumps together honest refusals to answer and substantively wrong predictions. Judges sometimes score refusals at 2 or 3 instead of 1, especially when the refusal recites related facts.
2. **Length-score correlation in C5.** Length and score correlate at r = 0.60 in the no-context baseline (C5), driven by hedging, adjacent-fact recitation, and disambiguation offers. The correlation is near-zero in Spec-containing and facts-containing conditions.

Direction of bias. Both effects raise C5 baseline scores more than they raise Spec-condition scores. The true Spec-vs-baseline gap is therefore likely *larger* than the +0.89 mean lift reported in §4, not smaller. The full audit (per-judge strictness, per-response-model abstention behavior, memory-system effect on abstention) is reported in §4.6.7; the analysis plan is left intact rather than recomputed under a modified rubric.

The class-level LLM-as-judge limitation that this methodology cannot fully address is treated in §6.2.[30]

3.4 Subjects

We test 14 subjects, all historical figures with public-domain autobiographies or memoirs. Subjects were selected across a range of time periods, source-text lengths, and geographic origins to avoid the study sitting on any single type of source material. All source corpora are English or English-translated and are available on Project Gutenberg or comparable public-domain archives. Because frontier language models train on large public-text corpora, some level of pretraining exposure to each subject's writing is likely.

#	Subject	Source	Words	Origin	Era
1	Philip Gilbert Hamerton	Project Gutenberg #8536	25,231	England	19th c.

[30] Raw per-judge judgments are in the public repository at `results/global_<subject>/*_judgments_<judge>.json` (and `judgments_v2.json` for the merged v2 set) for the 13 global subjects, `results/hamerton/*_judgments_<judge>.json` for Hamerton, and `results/franklin/*_judgments.json` plus `results/franklin_legacy_20260411/analysis/*_judgments.json` for Franklin. Memory-system per-judge judgments live at `results/global_<subject>/<system>_judgments_<judge>.json` (controlled) and `results/global_<subject>/<system>_fullpipeline_judgments_<judge>.json` (native) in the same flat per-subject directory.

#	Subject	Source	Words	Origin	Era
2	Elizabeth Keckley	Project Gutenberg #24968	58,742	United States	19th c.
3	Sunity Devee	Project Gutenberg #57175	67,379	India	19th–20th c.
4	Zitkala-Ša	Project Gutenberg #10376	35,328	United States (Yankton Dakota)	19th–20th c.
5	Olaudah Equiano	Project Gutenberg #15399	85,660	West Africa (Igbo) / Britain	18th c.
6	Mary Seacole	Project Gutenberg #23031	62,467	Jamaica	19th c.
7	Fukuzawa Yukichi	Internet Archive	139,088	Japan	19th–20th c.
8	Bābur	Project Gutenberg #44608	422,772	Central Asia (Fergana)	15th–16th c.
9	Yung Wing	Project Gutenberg #54635	66,459	China	19th–20th c.
10	Benvenuto Cellini	Project Gutenberg #4028	190,390	Italy (Florence)	16th c.
11	Bernal Díaz del Castillo	Project Gutenberg #32474	187,315	Spain (Castile)	15th–16th c.
12	Georg Ebers	Project Gutenberg #5599	96,174	Germany	19th c.
13	Jean-Jacques Rousseau	Project Gutenberg #3913	278,120	Switzerland (Geneva)	18th c.
14	Saint Augustine	Project Gutenberg #3296	114,873	North Africa (Numidia)	4th–5th c.

Constraints on the generalizability of the 14-subject sample (language, era, cultural framing, Project Gutenberg curation bias, individual self-presentation in autobiography) are discussed in §6.1.

3.4.1 Pretraining-coverage variance

Pretraining coverage of a specific person varies widely across the 14 main-study subjects, even within a sample whose autobiographies are of comparable provenance. We use the C5 baseline

score (no-context prediction accuracy on the 1-5 rubric defined in §3.3) as the observable proxy for pretraining coverage.

Baseline band	Subjects	Count
$\leq$ 2.0 (low-baseline)	Sunity Devee, Ebers, Hamerton, Fukuzawa, Seacole, Bernal Díaz, Keckley, Yung Wing, Bābur	9
2.0–3.0 (mid-baseline)	Cellini, Zitkala-Ša, Rousseau, Augustine, Equiano	5
$>$ 3.0 (high-baseline)	Benjamin Franklin (known-figure control, not in main study)	1

The higher a subject sits on this distribution (e.g., Equiano, Zitkala-Ša at the upper end of the main study), the better the model already knows them from pretraining; conversely, the lower they sit (Sunity Devee, Ebers), the less the model knows. Nine of the 14 main-study subjects fall into the low-baseline band; five into the mid-baseline band.

Franklin as a known-figure control. Benjamin Franklin (Project Gutenberg #20203) is included as a high-baseline reference point. Franklin's *Autobiography* is one of the most widely available and frequently cited autobiographies in American public-domain literature, and the model's baseline score on Franklin (3.77 on the 5-judge primary panel; 4.10 on Haiku alone) is consistent with substantial pretraining coverage of both the person and the specific text.[31] Franklin is used to anchor what the high-baseline end of the spectrum looks like (§4.6.4 develops the high-baseline behavior), not as a subject whose pretraining coverage is a design target of the Behavioral Specification.

Cross-provider variance. No-context baseline scores on the same subject vary by 1–2 points across response models. Different providers know different amounts about the same historical figure (Tier 2 cross-provider data in §4.6.1). The pretraining-coverage variance documented in this section is a property of each model family, not a property of the subject alone. Models disagree on who they have been trained on, by roughly the same magnitude that the specification lifts prediction.

§4.1 develops this distribution into the cross-subject gradient.

3.5 Question battery formation

Each subject's behavioral prediction battery is generated by a backward-design process: an LLM reads a passage from the held-out half of the corpus, writes a question whose answer is the behavioral pattern implicit in the passage, and avoids naming any detail unique to the passage itself. The question can be attempted from training-text patterns alone; the verbatim held-out passage is the ground truth for scoring.

The procedure, in order:

[31] Franklin's legacy run pre-dates the gemini_flash / gemini_pro split, so only one Gemini judge is present in the legacy data. The per-judge range cited here is the min and max across all available judges, rather than a clean 7-judge panel mean.

1. **Split the corpus.** Each subject's source text is divided by chapter into `training.txt` (50% of the corpus) and `heldout.txt` (50%). Held-out text is never shown to a response model.

2. **Sample from held-out.** A sliding window moves across the held-out text in four batches of ten questions each, using 5,000-character windows as local context.

3. **Backward-design question generation.** Claude Haiku 4.5 (temperature 0) reads each held-out window and writes a question whose answer requires the subject's behavioral patterns observable in the training half. For each window, Haiku:

   - Extracts a verbatim ground-truth span from the held-out window.
   - Avoids named-entity or specific-date leakage in the question stem.
   - Targets one of 10 fixed behavioral-prediction categories: decisions, values, relationships, conflict, learning, risk, creativity, stress, career, and change-over-time.

4. **Supplementary tiers (not scored in §4).** Beyond the behavioral-prediction tier, each battery additionally contains supplementary question tiers that test different competencies:

   - **Recall** tests whether a system can retrieve specific factual content.
   - **Adversarial-abstention** presents scenarios the source corpus does not cover, testing whether a system correctly refuses to predict rather than fabricates.

   These supplementary tiers are included in the battery files for future analysis but do not enter the §4 main results.

5. **Dedup and freeze.** Deduplication on lowercased question text, cap at target counts per category, MD5 checksum of the final battery. Downstream response and judgment files are invalidated if the battery checksum changes.

Each main-study subject receives 39 behavioral prediction questions; the 14 main-study batteries total 546 questions.[32] Each battery covers 8 to 10 of the 10 categories. Definitions, example questions, and per-subject distributions are in Appendix B.1 and B.2.

Within the behavioral-prediction tier, individual questions vary in interpretive demand: some can be answered from facts alone, others require applying behavioral patterns to novel scenarios. The Spec concentrates its benefit on the interpretation-heavy slice; full decomposition in §4.4.3.

Leakage audit. We empirically checked the no-leakage principle by searching every behavioral-prediction question for any sequence of seven or more consecutive words that appears verbatim in that subject's held-out corpus. Across the 14 main-study subjects (546 questions), zero questions leak (true zero, not a rounded value).[33]

One false-premise outlier. Zitkala-Ša Q18 presupposes she faced execution, which she did not; the backward-design prompt produced a malformed item. This is 1 of 586 questions (0.17%), does not affect any aggregate claim, and points to the broader limitation that automated batteries need a human-reviewed quality gate at scale (§6.2).

Battery files and the leakage-audit script are in the public repository.[34]

[32] Franklin's 40-question legacy battery (high-baseline reference, §3.4.1) brings the data-structural total to 586 questions across 15 subjects.

[33] Franklin's 40-question legacy battery, included as a high-baseline reference but not part of the main study, has 2 leaking questions (Q49, Q56). Both predate the backward-design constraint and were hand-authored. Aggregate across the full 586-question pool: 2 leaks (0.34%).

[34] Per-subject batteries at `results/global_<subject>/battery_v2.json` (13 global subjects); Hamer-

3.5.1 Circularity controls

The pipeline and the batteries both use Anthropic models for several roles (Haiku for extraction and battery generation, Sonnet for authoring, Opus for composition, Haiku as the primary response model, plus Sonnet and Opus on the judge panel). To rule out a within-Anthropic frontier-model artifact we ran two independent controls; full per-subject results for both are reported in §4.6.1.

Control 1: Independent battery regeneration. We re-generated batteries for all 13 global subjects with GPT-5.4 using the identical backward-design prompt and observed only minor categorical-emphasis differences (GPT-5.4 leaned toward risk and change-over-time questions; Haiku toward values and decisions). The methodology constrains the output more than the generating model does. Per-subject detail in §4.6.1.

Control 2: Non-Anthropic response chain. We re-ran the core conditions[35] on three subjects (Ebers, Yung Wing, Zitkala-Ša) using two non-Anthropic response models (Claude Sonnet 4.6 and Google Gemini 2.5 Pro) reading the GPT-5.4-generated batteries. The combination tests whether the Spec effect survives when both the response model and the battery-generation model sit outside the Anthropic family. Full results in §4.6.1.

A broader LLM-as-judge circularity (whether any LLM panel might systematically favor LLM-produced outputs over human-written alternatives) is not addressed by these controls; it is discussed as an open limitation in §6.2.[36]

3.6 Response models

Tier 1 (main study): Claude Haiku 4.5 as the primary response model, run across all 14 subjects and every condition in the main matrix. Haiku was chosen as a deliberately weaker baseline: a Spec effect that registers on a relatively weak response model is harder to attribute to the model's pretraining alone, which gives a conservative readout of effect direction. §4.6.1 Tier 2 cross-provider probe tests whether the direction reproduces on stronger response models.

Tier 2 (cross-provider response generation). Same configuration as Control 2 (§3.5.1): Claude Sonnet 4.6 and Google Gemini 2.5 Pro as response models on three subjects (Ebers, Yung Wing, Zitkala-Ša) reading the GPT-5.4-regenerated batteries. Full results in §4.6.1.

Call-time parameters. All response models are called with `temperature=0` and `max_tokens=1024`.

Prompt schema. A single shared prompt is used across every condition. The system message frames the task as behavioral prediction of a specific person; the user message is the question plus whichever context inputs the condition specifies (§3.2). Nothing about the prompt changes per condition beyond the injected context block.

```
System: You are predicting how <subject> would respond to a specific
        question about their behavior, values, or reasoning. Answer
        in <subject>'s voice, grounded in their demonstrated patterns.
```

ton and Franklin legacy at `data/<subject>/battery.json`; GPT-5.4-regenerated batteries (used in the §3.5.1 circularity control) at `results/global_<subject>/battery_gpt54.json`; leakage-audit script at `scripts/_verify_battery_leakage.py`.

[35] C5 no-context baseline, C2a Spec alone, C4a facts-plus-Spec, C2c wrong-Spec control. Full condition definitions in §3.2.

[36] Raw battery regeneration data at `results/global_<subject>/battery_gpt54.json` for all 13 global subjects; Tier 2 response and judgment files for the three Control 2 subjects in the same per-subject directories.

```
User:   <context block, one of: empty (C5), Spec (C2a), wrong-Spec (C2c),
         facts (C4), facts + Spec (C4a), corpus (C8), corpus + Spec
         (C9), or retrieval ± Spec (C1 / C3)>

        Question: <question text>
```

The prompt is deliberately uniform and deliberately faithfulness-oriented. The system message asks the model to ground its answer in the subject's demonstrated patterns and to answer in the subject's voice; we chose this framing because the study tests whether the served context lets the model do exactly that. No instruction tells the model to abstain, hedge, or commit; the model's natural refusal-or-commitment pattern given a specific context is itself part of the phenomenon the study tests, and §4.3 reports the hedging-rate shift across conditions as a substantive finding rather than a behavior to suppress.

Two effects of the faithfulness framing to acknowledge. Asking for the subject's voice may benefit systems that surface verbatim training-corpus passages over systems that surface compressed interpretive structure; Letta's stateful-agent path is one such system, examined post-hoc in §4.5 and Appendix G. Asking the model to ground in demonstrated patterns may also push the model toward abstention when the served context underdetermines a confident answer; that pattern is visible in the Spec-induced refusal cases discussed in §3.3.6 and §4.4.3. We surface these as real interactions between the prompt and the conditions rather than as bugs in the prompt. The framing is identical across every condition, so its priming is a constant; the differential effect the study measures (correct Spec vs no Spec, correct Spec vs wrong-Spec) is what the conditions vary against that constant. A neutral-prompt robustness check, without the "voice" and "grounded in demonstrated patterns" framing, is flagged as future work in §7.

What the conditions vary, given a fixed framing, is where the patterns come from.

Condition	Pattern source served to the model
C5 (no context)	The model's pretraining alone
C2a (Spec only)	Pre-extracted patterns served as the Behavioral Specification
C4 / C8 (facts / corpus)	Raw material; the model must identify patterns at runtime
C4a / C9 (facts + Spec / corpus + Spec)	Pre-extracted patterns alongside raw material
C2c (wrong-Spec)	Pre-extracted patterns from a different person

The empirical evidence that the prompt is not the lift mechanism (the C2a vs. C4 / C8 comparison and the C2c wrong-Spec control) is reported in §4.3; we do not preview the numbers here.

Exact model identifiers, full prompt text, and Tier 2 invocation parameters are in Appendix C.[37]

[37] Run scripts at `scripts/run_global_subjects.py`, `scripts/run_full_study.py`, `scripts/run_multimodel_responses.py`. Raw response files at `results/global_<subject>/results_v2.json` (13 globals), `results/hamerton/results.json`, `results/franklin/fullstack_haiku.json`, and `results/_tier2/` for Tier 2 runs.

3.7 Pipeline for the Behavioral Specification

The Behavioral Specification is a structured document encoding a person's behavioral patterns across three interpretive layers (anchors, core, predictions) plus a composed unified brief; total size per subject is approximately 7,000 tokens (~5,000 words, about the length of a short magazine article). It is produced by the Base Layer pipeline (this paper's open-source reference implementation), which transforms raw source text in five steps: import, extract, embed, author, and compose. Each step is a single script backed by a single model choice.

Step	Input	Tool / model	Output
1. Import	ChatGPT / Claude exports, journals, plain text, directories	`import_conversations.py`	A local database holding the cleaned, de-duplicated source text
2. Extract	Canonical source text	`extract_facts.py`, Claude Haiku 4.5, 46-predicate vocabulary	Behavioral patterns extracted as short structured statements (e.g., "avoids confrontation," "values craft over speed"), with bookkeeping operations to add new patterns, update existing ones, delete contradicted ones, or skip duplicates
3. Embed	All imported message text	`embed.py`, `all-MiniLM-L6-v2`, ChromaDB	A searchable index of source passages. Combined with the source-message IDs each extracted fact carries, this lets any claim in the final specification be traced back to the passages that support it

Step	Input	Tool / model	Output
4. Author	Extracted facts	`author_layers.py`, Claude Sonnet 4.6	Three interpretive layers as markdown (anchors, core, predictions); see body below for layer-by-layer examples. Each layer is produced from facts alone, not from prior layer output. Each layer prompt includes a domain guard that prevents topic skew (ablation-validated in prior pilot work).
5. Compose	The three authored layers (plus a sample of identity-tier facts as supplementary context)	`agent_pipeline.py`, Claude Opus 4.6	Unified behavioral brief in flowing prose; see body below for what the brief contains and why it exists

The artifact served as context in experimental conditions is the three authored layers concatenated with the composed brief, not the brief alone.

The extract step constrains output through a fixed vocabulary of 46 behavioral predicates (examples: `avoids`, `repeatedly engages in`, `refuses to`, `values`, `fears`, `has experienced`). The full predicate list is in Appendix A. The vocabulary is human-curated and was validated across 50+ pilot subjects before being frozen for the study. The constrained vocabulary is the main lever the pipeline uses to push extraction away from biographical facts ("his father was violent") and toward behavioral patterns ("evaluates authority figures on dual criteria of virtue and failure").

The three authored layers have distinct jobs. Each layer has a characteristic format; examples below are drawn from three different subject specifications (Sunity Devee, Bernal Díaz, and Augustine).

Anchors encode the subject's load-bearing axioms in numbered form (A1, A2, ...), each with an activation condition and a false-positive warning. Example from Sunity Devee:

> *A1. DIVINE PRIMACY. All events, decisions, and outcomes are understood as expressions of divine providence first; social, political, or material explanations are secondary framings applied afterward. Active when: life decisions, loss, political events, marriage, reform, or any outcome described as fortunate or unfortunate are discussed.*

Core captures values, beliefs, and self-view in flowing prose. It is the layer that reads most like an essay about the person. Example from Bernal Díaz:

> *They reason from direct witness and lived participation, treating firsthand account as the only reliable foundation for any claim. When evidence is secondhand or reconstructed, they flag it rather than smooth it over. They distrust narrators who were absent from*

the events they describe, and this distrust shapes how they evaluate any source placed before them.

Predictions are explicit behavioral predicates (P1, P2, ...) with detection criteria, directives, and false-positive warnings. Example from Augustine:

> *P1. CONFESSION BEFORE CONCLUSION. When asked to account for a past failure, does not defend, minimize, or contextualize first. Moves immediately into detailed reconstruction of the failure, naming the pleasure or pride taken in the transgression before offering resolution. Detection: a relational conflict where the subject was at fault, a professional misjudgment, a moment of intellectual dishonesty. Directive: hold space for the full weight of the confession; resist premature resolution. False positive: not active when the subject is analyzing someone else's failure.*

The unified brief. The compose step integrates the three authored layers into a continuous prose synthesis in the third person, similar in length to a short profile of the subject. The brief serves a dual purpose: a coherent first pass for human readers, and an integration step that implicitly weaves the three layers together (layered files alone do not require this integration; internal testing suggests the integration changes how a model uses the specification). A formal ablation isolating brief-with-layers vs. layers-only is flagged in §7.3 Specification design and composition.

Total pipeline cost is approximately $1 per subject (estimated from per-step API token counts) to process a 50,000- to 150,000-word autobiography end to end. Pipeline code, the full predicate vocabulary, and example specifications for all 14 study subjects are available in the public repository (see §8 Data, code, and reproducibility).

---

4. Results

Across 14 historical subjects, adding a Behavioral Specification (a short structured document describing how a specific person reasons and behaves) measurably improves how accurately a language model represents that person's behavioral patterns. We measure this with a battery of behavioral prediction questions based on held-out ground-truth text from each subject's publicly available autobiography. We score each prediction on a 1-to-5 rubric where a whole-point shift marks a categorical change in how the response aligns with the subject's documented behavior.

On the 9 low-baseline subjects (those the model does not know well), the specification produces a mean per-subject increase of **+0.89 points** and lifts individual responses by one category or more on **55.0% of questions**. The specification's added value on top of other context types (facts, raw corpus, or memory-system retrieval) concentrates on interpretation-heavy questions; on factual-recall questions, retrieval alone is often sufficient and the specification adds little or actively degrades the response. On high-baseline subjects (those the model does know well, such as Benjamin Franklin), the specification adds little or mildly hurts across conditions. Control conditions, statistical robustness checks, and sensitivity analyses confirm that the specification categorically shifts how a language model responds, increasing its ability to hold an accurate representation of the subject beyond what pure fact-based retrieval can supply.

The seven parts of §4 establish this picture in detail:

- **§4.1. The cross-subject gradient.** The primary result, across 14 subjects.

- **§4.2. Compression: structure vs. raw text.** Is the effect about structure or about information volume?
- **§4.3. Mechanism: Content, Not Format.** Does the content of the correct specification drive the effect, or does any structured prompt?
- **§4.4. Memory-system composition.** Does the specification layer on top of existing commercial memory systems? Where does it help or hurt at the per-question level (§4.4.3 common mechanisms, §4.4.4 cross-system Keckley case)?
- **§4.5. Exploratory case study (Letta stateful-agent).** Brief summary in body; full case study in Appendix G. Post-hoc N=3 comparison; not a headline finding.
- **§4.6. Robustness and sensitivity.** Cross-provider response generation, judge-panel sensitivity, battery composition sensitivity, wrong-Spec derangement protocol sensitivity, retrieval-overlap sensitivity, rubric-handling limitations from a post-hoc validity audit, and what these checks do not address. (The high-baseline end of the gradient through the Franklin reference is in §4.1.2.)
- **§4.7. Summary and bridge to discussion.** A one-paragraph synthesis of what §4 established, framing the transition into §5.

Every number in §4 uses the 5-judge primary aggregate defined in §3.3.3 (Haiku 4.5, Sonnet 4.6, Opus 4.6, GPT-4o, GPT-5.4). The 7-judge sensitivity check (adding Gemini 2.5 Flash and Gemini 2.5 Pro) is reported in §4.6. Score deltas are read through the anchor-crossing rule from §3.3.1: a delta that crosses a rubric integer anchor is a stronger claim than one that stays inside a single anchor band.

4.1 The cross-subject gradient and its per-question mechanism

Hypotheses tested in this section (from §1.2): H1. Adding the specification improves prediction. H2. The effect is inversely proportional to the response model's pretraining coverage. Corollary to H2: on high-baseline subjects, the specification does not add value and mildly interferes.

The cross-subject gradient. The less the model already knows about a subject from pretraining, the more the Behavioral Specification improves the model's representational accuracy of that subject. It operates as an interpretive layer over facts and retrieved context, not a replacement for them. On the 9 subjects whose pretraining baseline sits at or below 2.0 on the 1-5 rubric (the population of relevance from §3.4.1), adding the Spec consistently improves prediction: every one of the 9 improves over no-context baseline (mean $\Delta = \mathbf{+0.71}$ for Spec alone, **+0.89** for facts + Spec); none declines. Adding the Spec on top of all extracted facts (C4), raw corpus (C8), or memory-system retrieval produces additional aggregate gains that are smaller in magnitude than the Spec-vs-baseline lift (detail in §4.2 and §4.4). Spec alone does not score higher than facts alone or raw corpus alone; the Spec's value is in the layering.

Adding a Behavioral Specification changes the category of answer the response model produces. Of the 351 individual responses in the low-baseline slice, **55.0% crossed at least one rubric integer anchor upward when the specification was added**. Multi-anchor jumps of two or more bands (e.g., 1→3, 2→4) appear in 18% of low-baseline questions on the Spec conditions, with about 6% being extreme jumps of three or more bands (e.g., 1→4, 2→5, 1→5). These extreme jumps concentrate on interpretation-heavy questions: the no-context response refuses or stays

generic, and the specification supplies the behavioral pattern the model could not retrieve from training data. The response model's answer moved from one category of response to a qualitatively different category. These are the multi-anchor jumps at the margin the aggregate mean understates.

The Spec as a leveler. The Spec levels prediction quality across subjects: regardless of how much the model knew about each subject going in, every subject ends up at roughly the same place on the rubric (per-subject mean facts + Spec score = **2.44** across all 14 subjects), clustering tightly in the 2.0–2.7 band.[38] The lift the Spec produces is therefore largest on subjects whose baseline starts low, smallest on subjects whose baseline already approaches the quality the Spec produces. **Most users sit in the low-baseline category, the population of relevance for AI personalization (§1.4, §5.2):** AI users whose private reasoning is not in any training corpus fall at or near the rubric floor by construction, so the lift is largest exactly where the use case is most common. The Spec works for any subject regardless of pretraining coverage and is therefore portable across the long tail of users.

Transition	% of responses	Description
$1 \rightarrow 2$	**33.3%**	Refusal or off-base → generic engagement with the question
$1 \rightarrow 3$	12.3%	Refusal → partially-aligned prediction
$1 \rightarrow 4$	4.8%	Refusal → substantively-aligned prediction
$1 \rightarrow 5$	0.9%	Refusal → fully matches the held-out pattern
$2 \rightarrow 3$	2.0%	Generic → subject-specific
$2 \rightarrow 4$	0.3%	Generic → substantively-aligned
$3 \rightarrow 4$	1.4%	Partially → substantively-aligned
No upward crossing	38.2%	Delta stayed inside a single anchor band
Downward crossing	6.8%	Specification hurt the response

One of every three low-baseline responses moves from "cannot engage" to actual engagement. Another one in five makes a larger jump. Only one response in fifteen gets worse. Worked examples of these transitions appear below (Examples A, B, and C) and in §3.3.1 (multi-anchor crossings) and §4.1.1 (Seacole Q2 across condition bands).

Three representative examples below show distinct mechanisms by which the specification improves the response.[39] Hedge reduction is common but not the only mechanism. The specification also corrects directionally-wrong predictions and enables interpretive inference from character patterns when retrieved facts are insufficient.

[38] Two regression slopes capture the gradient. The level slope of C4a on C5 is essentially flat (+0.04, $R^2 = 0.008$): the Spec produces roughly uniform quality regardless of where each subject's baseline starts. The change-score slope of Δ_C4a on C5 is strongly negative (−0.96), but most of that magnitude is mechanical: slope_Δ = slope_level − 1 by identity, so a flat level slope automatically produces a near-−1.0 change-score slope. Full technical sensitivity detail in §4.6.3.

[39] Raw per-judge scores and full response text for the three examples at `results/global_<subject>/results.json` and `results/global_<subject>/judgments_*.json` (Examples A, B, and C from `results/global_fukuzawa/`, `results/global_bernal_diaz/`, and `results/global_seacole/`). C4 = all extracted facts only; C4a = all extracted facts + Spec.

Example A. Facts → Facts + Spec: voice-matched argument from character pattern

Subject: Yukichi Fukuzawa. **Q35:** *"Would Fukuzawa's reasoning for keeping students from war be based on moral principle or practical concern?"*

Held-out: *"For this kind of war, if they wanted figureheads in their ranks, they could as well hire farmers from the provinces. My students were too precious."*

C4 facts only (mean 1.00): *"Fukuzawa's reasoning would likely be* ***practical concern*** *rather than moral principle... his demonstrated pragmatism... his attitude toward violence... a personal squeamishness rather than a principled opposition... his political detachment... students were more valuable to Japan's future as scholars than as soldiers — a practical calculation rather than a moral one."*

C4a facts + Spec (mean 4.20): *"****A2 (Utility Gate)*** *is his load-bearing axiom here. He reasons from what a thing does before defending what it is. When advising students to avoid war, he would lead with:* ***Capability preservation:*** *'You cannot advance your learning if you are dead or conscripted.'* ***Opportunity cost:*** *'The war will still exist in six months; your education window will not.'* ***Competitive advantage:*** *'Those who master Western knowledge during this chaos will position themselves when the dust settles.' "*

Mechanism: C4 produces correct-direction third-person analysis ("Fukuzawa would think practically") but never lets Fukuzawa speak; the panel scored it 1.00, treating the meta-analysis as off-target. C4a applies the A2 (Utility Gate) axiom to put first-person voice-matched arguments in Fukuzawa's mouth (capability preservation, opportunity cost, competitive advantage). The held-out's specific "hire farmers as figureheads" angle is not reached, but the *way Fukuzawa would reason* is captured directly. 1→4 jump (1.00 → 4.20), the Spec's added value isolated against an all-facts baseline. **Mechanism: voice-matched argument from character pattern.**

Example B. Facts → Facts + Spec: directional correction

Subject: Bernal Díaz del Castillo. **Q16:** *"When the commander is offered physical assistance during a strenuous activity, what behavior would the author expect from him?"*

Held-out: *"There were 114 steps to the summit, and, as they feared that Cortes would experience the same fatigue in mounting as Motecusuma had, they were going to assist him by taking hold of his arms. Cortes, however, refused."*

C4 facts only (mean 2.00): *"Based on the facts provided about Bernal Díaz del Castillo and his observations of Cortes, the author would expect the commander to* ***accept*** *and participate in the assistance while maintaining leadership and setting an example."*

C4a facts + Spec (mean 4.80): *"Based on the behavioral specification, when the commander is offered physical assistance during a strenuous activity, the author would expect Cortes to* ***refuse*** *the assistance and perform the labor himself, particularly if it serves a symbolic or morale-signaling purpose."*

Mechanism: all facts alone produced a confident but directionally wrong prediction (commander accepts help). The specification corrected the prediction to match the held-

out (Cortes refuses), encoding Cortes's pattern of physical self-reliance and performative leadership over the model's generic "good leaders accept help" default. Directional correction.

Example C. Facts → Facts + Spec: abstention becomes near-perfect inference

Subject: Mary Seacole. **Q2:** *"How does Mary Seacole typically respond when a delirious patient mistakes her for a family member?"*

Held-out: *"Poor fellow! I could not undeceive him. I think the fancy happily caused by the touch of a woman's hand soothed his dying hour."*

C4 facts only (mean 2.80): *"Based on the complete set of facts provided, there is **no specific information** about how Mary Seacole typically responds when a delirious patient mistakes her for a family member. While the facts document many aspects of Mary Seacole's nursing care and compassion, there is **no explicit account** of her reactions to delirious patients confusing her identity with their relatives."*

C4a facts + Spec (mean 5.00): *"Based on the behavioral specification and known facts, Mary Seacole's response to this situation would follow a consistent pattern. Primary response: compassionate acceptance. She would **not correct the patient** in the moment. Instead, she would **lean into the mistaken identity, sustain the fiction with tenderness**, and prioritize emotional relief over factual accuracy. This aligns with A4 (Proximate Obligation) and A8 (Sympathetic Realism)."*

Mechanism: clean abstention under all facts alone (the model identifies the gap and refuses to invent) lifts to specific accurate inference under all facts + Spec. **2.80 → 5.00 jump, crossing three integer anchors upward.** The specification enabled generalization from Seacole's compassionate-caregiving pattern (documented in the facts but not explicitly mapped to this scenario) to the specific untested situation. Interpretive inference beyond what the all-facts dump alone provides.[40]

Two statistical confirmations support the directional claim. The **Wilcoxon signed-rank test** across all 14 main-study subjects confirms that the specification's lift over baseline is real and not due to chance.[41] **Pairwise Spearman** ρ across the 5-judge primary panel (§3.3.4) confirms that the lift's direction is consistent across judges rather than dependent on any one judge's scoring.[42]

---

[40]The judge panel scored the C4 abstention at 2.80, not 1.00. Judges treat honest abstentions as partial engagement (~2.5–3.0); they sometimes also penalize Spec-induced honest abstentions where the Spec appropriately declined to invent detail (§4.4.4's Keckley Q21). The rubric does not cleanly distinguish abstention from wrong prediction (§3.3.6); a differentiated rubric is flagged in §7.

[41]Wilcoxon signed-rank test results on the 14 main-study subjects: C5 vs. C2a $W = 10$ ($p = 0.005$); C5 vs. C4a $W = 11$ ($p = 0.007$).

[42]Pairwise Spearman ρ across the 5-judge primary panel runs 0.86 to 0.93. Regression of Δ_C4a on C5 baseline: slope $\mathbf{-0.96}$ [95% CI -1.24, -0.67], $R^2 = 0.82$, $p < 0.001$ ($p = 0.000009$), correlation $r = -0.90$. Subjects with positive Δ_C4a: 12 of 14; low-baseline subjects ($n = 9$) all positive; low-baseline mean Δ_C4a = $\mathbf{+0.89}$. Subject-level bootstrap (10,000 resamples) gives 95% CI $[-1.25, -0.74]$, excluding both zero and a 50%-attenuated effect; full bootstrap, joint regression, and permutation tests in §4.6.4. Full statistical detail in Appendix B.6.

Reading the gradient. Figure 4.1 plots each subject's no-context baseline (C5) against the lift the specification produces over that baseline (Δ_C4a). The slope is the core relationship: subjects with lower baselines see larger lifts; subjects with higher baselines see smaller or negative lifts. The 9 low-baseline subjects (C5 $\leq$ 2.0) cluster in the upper-left of the plot with positive lifts ranging from Bābur at **+0.25** (smallest lift) to Hamerton at **+1.51** (largest). Franklin sits in the lower-right at C5 = 3.77, Δ = −0.13: the high-baseline reference where the model already knows the subject from pretraining. The regression slope of **−0.96** captures this gradient: the lower the model's pretraining baseline on a subject, the larger the lift the specification produces, because the Spec produces a roughly constant facts + Spec quality near 2.44 regardless of baseline. The takeaway: the specification helps most where the model knows the subject least; once a subject crosses into the high-baseline band, the specification has no representational gap to fill.

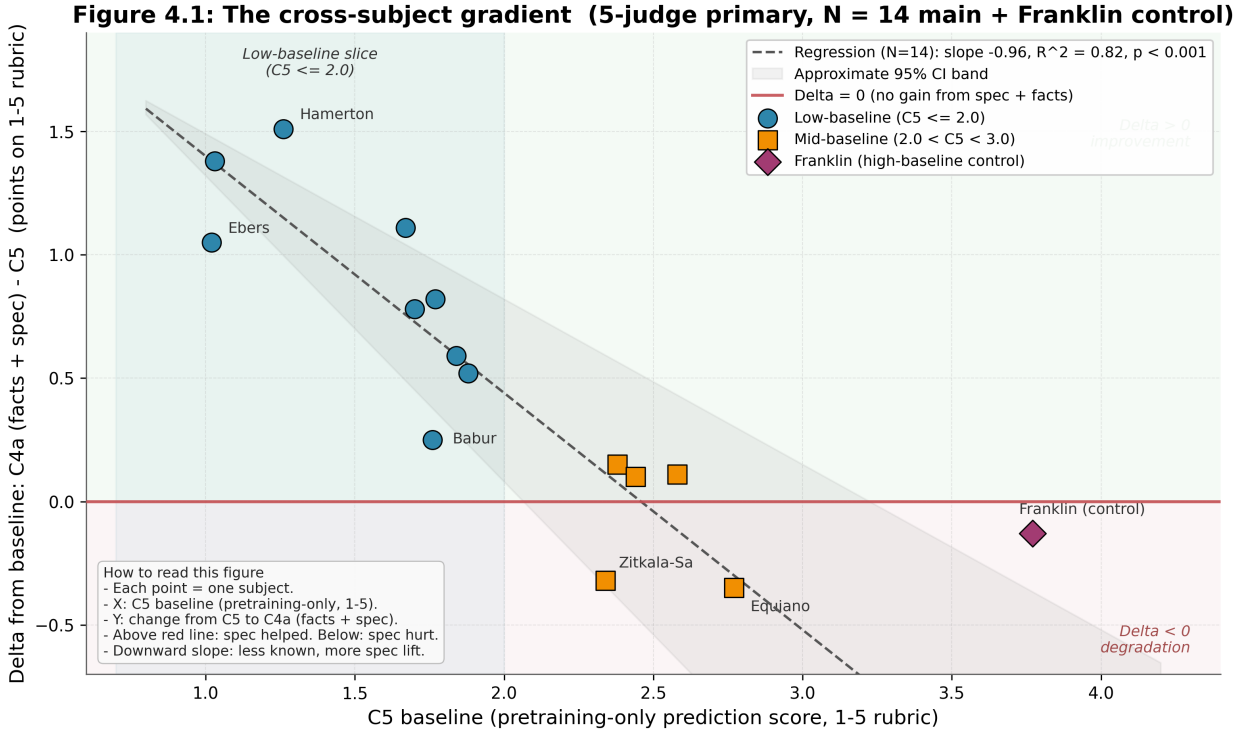


Figure 1: Figure 4.1: Cross-subject gradient. Each subject's no-context baseline (C5, x-axis) plotted against the specification lift (Δ_C4a, y-axis) for all 14 main-study subjects. Low-baseline subjects (C5 $\leq$ 2.0, the population of relevance) cluster in the upper-left with positive lifts ranging from Bābur (+0.25) to Hamerton (+1.51). Franklin (high-baseline reference, C5 = 3.77) sits in the lower-right with Δ = −0.13. Regression slope −0.96, $R^2 = 0.82$. (§4.1)

Per-subject results.

The table is ordered by baseline within each band. In the color-rendered PDF of the paper, the low-baseline rows are tinted green (the population of relevance), the mid-baseline rows are tinted yellow, and Franklin is tinted gray as the high-baseline reference. Figure 4.1 presents the same data as a scatter plot with the regression line.

Bands: **Low** (C5 $\leq$ 2.0, the population of relevance), **Mid** (2.0 < C5 < 3.0), **High** (Franklin, the

known-figure reference).

Band	Subject	C5 baseline	C4 facts	C2a Spec	C4a facts+Spec	Δ C4a−C5	Δ C4a−C4	Anchor
Low	Ebers	1.02	2.02	1.54	2.07	+1.05	+0.05	✓
Low	Sunity Devee	1.03	2.46	2.27	2.41	+1.38	−0.05	✓
Low	Hamerton	1.26	2.43	2.63	2.77	+1.51	+0.34	✓
Low	Fukuzawa	1.67	2.67	2.35	2.78	+1.11	+0.11	✓
Low	Bernal Díaz	1.70	2.41	2.27	2.48	+0.78	+0.07	partial
Low	Bābur	1.76	2.03	1.91	2.01	+0.25	−0.02	-
Low	Seacole	1.77	2.63	2.48	2.59	+0.82	−0.04	✓
Low	Keckley	1.84	2.39	2.43	2.44	+0.59	+0.05	-
Low	Yung Wing	1.88	2.13	2.22	2.40	+0.52	+0.27	-
Mid	Zitkala-Ša	2.34	2.10	2.03	2.02	−0.32	−0.08	-
Mid	Cellini	2.38	2.42	2.54	2.53	+0.15	+0.11	-
Mid	Rousseau	2.44	2.32	2.81	2.53	+0.10	+0.21	-
Mid	Augustine	2.58	2.56	2.48	2.70	+0.11	+0.14	-
Mid	Equiano	2.77	2.43	2.46	2.42	−0.35	−0.01	-
High	Franklin	3.77	—	3.37	3.65	−0.13	—	-

C4 (facts only) was generated and judged on the 9 low-baseline subjects and 5 mid-baseline subjects under the 5-judge primary panel; Franklin's C4 responses were generated but never scored under the 5-judge primary panel and remain dashed in the table.[43] The Δ C4a−C4 column shows what adding the specification contributes on top of facts alone. The Spec-on-facts increment is small and mixed in sign across both bands (low-baseline mean **+0.09**, mid-baseline mean **+0.07**), with most of the lift coming from the Spec-vs-baseline gap.

What each band is telling us.

- **Low-baseline (n = 9):** every subject improves. The slice is uniform. This is the population of relevance for real AI deployment.
- **Mid-baseline (n = 5):** 3 subjects improve, 2 decline. The model has enough pretraining footprint on these subjects that the specification competes with the model's own working model. The specification sometimes increases representational accuracy and sometimes does not.
- **Franklin (high-baseline reference):** both Spec-containing conditions score below baseline. The specification cannot add what the model already has.

Per-subject anchor-crossing distributions (ranging from 25.6% on Bābur to 74.4% on Sunity Devee)

[43] Mid-baseline C4 means recomputed from `results/global_<subject>/judgments_v2.json` 2026-05-07 against the canonical 5-judge primary aggregation (per-question 5-judge mean, then mean across questions): Zitkala-Ša 2.10, Cellini 2.42, Rousseau 2.32, Augustine 2.56, Equiano 2.43 (39 questions each). Earlier drafts of this table omitted these values; correction landed during §4.1 walk. Franklin C4 was scored only by the 2-judge legacy panel (Haiku + Gemini) and is not directly comparable to the 5-judge primary aggregate; remains a dashed cell.

and per-subject per-judge score matrices are in Appendix D.[44]

The aggregate gradient hides per-question structure. The specification produces large category-level shifts on a subset of questions (multi-anchor crossings, including band-5 endpoints reached from band-2 starts under cross-condition comparisons such as C4 → C4a) and minimal change on others. §4.1.1 decomposes this distribution and shows where the Spec's value concentrates. §4.2 takes the same gradient and asks whether the lift is about structure or about information volume, comparing the Spec against far larger raw-corpus context.[45]

4.1.1 Per-question baseline engagement and the worked rubric example

Across 546 questions on the 14 main-study subjects (5-judge primary panel), the no-context baseline (C5) splits into two clusters with a thin middle. **Roughly 41% of questions return a refusal or misalignment** (rubric score = 1.00; the model declines, names the wrong person, or lands far outside the question). **Roughly 21% return an answer specifically about the named subject** that engages with the question (rubric score ≥ 3.0). The band between is sparse. **The Spec moves refusals and misalignments into substantive predictions on 94.2% of bottom-cluster questions.** On the top cluster, where the baseline already produced a substantive answer, **the Spec did not help on roughly 79% of questions** and reduced the score on average.

The two findings together describe the per-question structure underlying the cross-subject gradient in §4.1. Where the baseline knows nothing about the subject, the Spec supplies the interpretive frame the baseline lacks. Where the baseline already engages with the subject, the Spec adds structure that does not improve the answer and sometimes reduces it. The response model was given a grounded-prediction prompt with no instruction to abstain and no cost to producing a confident wrong answer beyond a low judge score; it declined on more than 40% of questions. Whatever logic providers ship to allow abstention or refusal to answer takes precedence over the prompt's instruction to predict.

A score of 1.00 means the model failed to produce a usable prediction about the named subject (full definition in §3.3.1, "What a 1 means and does not mean"). In about 93% of score-1.00 responses, the model explicitly declined to answer ("I don't have enough information about this person"). The remaining 7% are non-abstention failures: the model engaged with the question, but the engagement was categorically incorrect (wrong referent, off-base inference, or confusion with a different subject). Both modes are addressable by adding the Spec, through different mechanisms.[46]

The pattern across baseline bins (X = C5 mean across 5-judge primary panel; lift = C4a − C5):[47]

[44] Two potential confounds on the gradient slope (battery-question-type composition and Hamerton-leverage subset regression) are addressed in §4.6.3 as robustness checks; both leave the baseline gradient effect substantially intact.

[45] Held-out leakage audit on the 60 unique extreme-upward-jump cases at `docs/research/held_out_leakage_investigation_20260428.md`: 0 6-gram matches at C4a, severity rare; full taxonomy and headline-impact estimate in Appendix B.9.

[46] The §3.3 rubric scores both explicit abstention and non-abstention misalignment as 1.00; the 93/7 split is from a post-hoc regex pass classifying responses (`scripts/analyze_baseline_engagement.py`). 7% is a coarse upper bound; finer analysis of *when and why* the model picks confident-wrong over abstention is open future work in §7.

[47] Spearman ρ between baseline X and Spec lift = $-\mathbf{0.73}$ (n = 546, p ≈ 1.7×10^{-91}). Per-subject ρ negative for 14 of 14; 12 of 14 reach $p < 0.01$.

C5 baseline (X)	N	Share	Spec lift mean (± SD)	Positive on	What this bin shows
X = 1.00	225	41.2%	+1.32 (± 0.88)	94.2%	Where the Spec does the most lifting; converts abstentions and confident-wrong answers into substantive predictions.
1.00 < X < 2.00	110	20.1%	+0.66 (± 0.83)	78.2%	Marginal baseline; Spec adds discrimination to partial answers.
2.00 ≤ X < 3.00	95	17.4%	+0.04 (± 0.63)	39.0%	Roughly even; Spec neither helps nor hurts on average.
3.00 ≤ X < 4.00	82	15.0%	−0.47 (± 0.81)	25.6%	Spec hurts on average; baseline already engages substantively.
X ≥ 4.00	34	6.2%	−0.99 (± 0.78)	8.8%	Spec hurts most; baseline already produces a near-correct answer the Spec does not improve.

The same bimodality also appears within each subject's 39-question battery. Three

patterns:[48]

- **Floor-saturated, 2 of 14 subjects.** More than 90% of the 39 questions in the battery return a refusal or misalignment from the baseline. The per-subject mean shows almost no variance.
- **Engaged-skewed, 1 of 14 subjects.** Fewer than 10% of the 39 questions in the battery return a refusal or misalignment. The baseline produces an answer specifically about the subject on most questions.
- **Mixed, 11 of 14 subjects.** The battery contains questions at both the floor (refusal or misalignment) and in the substantive-engagement range. Some subjects in this group sit closer to floor-saturated, others closer to engaged-skewed. All the high-baseline subjects in the §4.1 gradient fall in this pattern. Even when the baseline knows the subject well enough to answer most questions substantively, several questions in the same battery still trigger floor-level refusals.

The aggregate per-subject means in §4.1 are the average across this internal split.

Worked rubric example: Seacole Q2 across condition bands. The bins above describe the per-question structure as aggregate counts. The example below shows what that structure looks like on a single question, traced across all five conditions. The Seacole question was used as Example C in §4.1; here it is presented across the full condition set so the band-by-band progression is visible. The example shows three things. First, what the X = 1.00 bin actually contains as model output (the C5 row). Second, what adding facts alone (C4), the Spec alone (C2a), and both together (C4a) each produce on the same question. Third, the cross-anchor interpretation rule of §3.3.1 in operation: a Spec-grounded answer moving from a band 1 refusal to a band 5 affirmation that names a specific behavioral pattern. Bands follow the rubric in §3.3.

Subject: Mary Seacole. **Q2:** *Based on Mary's character, would she comfort a delirious soldier who mistakes her for his wife?*

Held-out ground truth: *"Poor fellow! I could not undeceive him. I think the fancy happily caused by the touch of a woman's hand soothed his dying hour."*

[48] *Floor-saturated:* Sunity Devee (37 of 39 questions at X = 1.00) and Ebers (36 of 39). *Engaged-skewed:* Equiano (only 2 of 39 questions at X = 1.00; 5 of 39 at X $\geq$ 4.00). *Mixed:* the remaining 11 subjects, including Augustine (8 of 39 floor-questions despite a per-subject C5 mean of 2.58).

Condition (natural-language)	Code	Mean	Band	Excerpt (verbatim)
Wrong Spec (a Bābur Spec, mis-applied to Mary Seacole)	C2c	1.40	1	"Mary would **not sustain the deception**, but she would **not abandon the soldier either**. … She would correct him, gently but directly … allowing him to die in a false belief violates her epistemic obligation."
No-context baseline	C5	1.80	1	"I'd need to know which Mary you're referring to. There are many famous Marys in literature, history, and other works. Could you provide more context?"
All extracted facts only (no Spec)	C4	2.80	2	"Based on the extensive facts about Mary's character, **yes, she would absolutely comfort such a soldier** … Mary would likely gently accept the soldier's confusion while providing comfort…"

Condition (natural-language)	Code	Mean	Band	Excerpt (verbatim)
Spec only (no facts)	C2a	4.20	4	"Yes, she would comfort him. But not by correcting him. … She would *not* say 'I am your wife' or sustain the delusion. … Stay present. Touch him. Speak to the need beneath the confusion."
All facts + Spec (full pipeline)	C4a	5.00	5	"Yes, absolutely … She would **accept the mistake without correcting it**, not out of deception, but out of A8 (Sympathetic Realism). She understands that his delusion is a mercy, a refuge his mind has built to survive unbearable pain."

The example traces the *cross-anchor interpretation rule* of §3.3.1 on a single question. C5 declines for lack of referent; C2c (a Bābur Spec) is articulate but mis-applied. C4 with raw facts reaches a generic affirmative. C2a, working from the full Spec without retrieved facts, produces an internally articulate multi-dimensional answer about presence and tenderness but lands on a *less specific* prediction about whether Mary would correct him. C4a, with both facts and Spec, explicitly endorses non-correction; the held-out shows Mary chose not to undeceive him, and C4a matches that direction outright while C2a is less clear on it. Categorical movement from band 1 through band 5 on a single question is what the per-subject means in §4.1 aggregate.

Per-response-model abstention behavior is named in §3.3.6 and decomposed in §4.6.7 (Sonnet

over-credits abstention at roughly twice Haiku's rate). Memory-system retrieval inflates refusal scores at the condition level rather than via visible fact recitation, decomposed in §4.4.

4.1.2 The gradient at the high-baseline end (Franklin reference)

Franklin is not part of the N=14 main-study sample. He is treated as a known-figure reference to test whether the gradient pattern holds at the high-baseline end, where the model already has substantial pretraining representation of the subject. **Franklin confirms the gradient at the high end: where the baseline is already near the ceiling, the Spec produces a small negative effect rather than a positive lift.** Franklin's *Autobiography* is one of the most widely cited autobiographical works in American public-domain literature, and every current-generation LLM has substantial pretraining representation of both the person and the text.

Franklin sits a full anchor band above the main study's upper end. His no-context baseline (C5) is 3.77 on the 5-judge primary panel; the next-highest main-study subject, Equiano, sits at 2.77. Franklin's baseline lands above rubric anchor-3 ("right domain, wrong outcome") and nudges toward anchor 4 ("general direction correct").

Both Spec-containing conditions score below Franklin's baseline. Spec alone drops 0.40 points; the full pipeline (facts + Spec) drops 0.13. The aggregate is the average of substantial per-question heterogeneity, the same pattern documented in §4.1.1: Spec lift is positive on 15 of 40 questions and negative on 20 of 40. Big positive lifts cluster on behavioral-prediction questions about scenarios the *Autobiography* does not narrate verbatim (Q38, +1.80, 2-anchor crossing; Q22, +1.60, 2-anchor crossing). Big negative lifts cluster on questions where the model already had the pattern (Q43, C5 = 5.00 → C4a = 1.80). The likely interpretation: the Spec alone disrupts the model's pretraining-derived representation of Franklin where the model already had it; adding facts back provides an anchor and partially recovers the baseline performance.[49]

The exact mechanism is not isolated in this study; further investigation is flagged in §7.[50]

4.2 Compression: structure vs. raw text

Hypothesis tested in this section (H5 from §1.2): A compact specification achieves comparable behavioral-prediction performance to the full raw source corpus, at a fraction of the context size.

Even within the compression frame, the per-question mechanism from §4.1.1 holds: on particular questions, the interpretive lens outperforms raw facts or raw corpus, even when the model has access to the full source text.

Context improves prediction. On the 9 low-baseline subjects, every context condition increases the per-subject mean score by roughly one full rubric point over the no-context baseline.

[49] Franklin's facts-only condition (C4) was scored only by the 2-judge legacy panel (Haiku + Gemini), not the 5-judge primary panel; Franklin's C4 cell is dashed in §4.1 Table 4.1 for that reason. The C4-vs-C4a comparison that isolates Spec contribution from facts contribution is therefore not available for Franklin. The interpretation in the paragraph above is theoretical and not isolated by an additional control.

[50] Raw per-subject Franklin data at `results/franklin_legacy_20260411/`. Per-question lift analysis from `scripts/analyze_baseline_engagement.py` applied to Franklin data 2026-05-07. Numerical claims (3.77 baseline, 0.40 / 0.13 drops, Q38 +1.80, Q22 +1.60, 15-of-40 positive count) to be added to `scripts/recompute_paper_numbers.py` per launch-blocking mechanistic-check infrastructure (§7).

Condition	Context served (approx. tokens, low-baseline mean)	Mean (low-baseline, n=9)	Δ from C5
C5	none (baseline)	1.52	0.00
C2a	Spec only (~7K)	2.23	+0.71
C4	all facts only (~10K)	2.35	+0.83
C8	raw corpus only (~163K mean; range 33K–549K)	2.45	+0.93
C4a	all facts + Spec (~17K)	2.45	+0.93
C9	corpus + Spec (~170K mean)	2.50	+0.98

The AI does not need much context to move from refusal-and-off-base to engaged subject-specific prediction. It needs *some* context.

The compact specification recovers 76% of the corpus's predictive lift at 23× less context. On the 9 low-baseline subjects, Spec alone (~7K tokens) produces +0.71 points of lift over baseline; the full raw corpus (~163K tokens mean, range 33K–549K) produces +0.93 points. The interpretive layer drives the result; the volume of context served does not.

The compression story holds across other context-size comparisons:

- **Spec vs all facts:** Spec produces 86% of the facts-only lift (+0.71 vs +0.83) at 30% less context.
- **Facts + Spec vs corpus alone:** the structured 17K-token package matches the raw 163K-token corpus's lift (+0.93 vs +0.93), at ~10× less context.
- **Facts + Spec vs corpus + Spec:** adding the entire raw corpus on top of facts + Spec produces less than 7% additional lift over baseline (+0.93 → +0.98), at 10× more context.

The behaviorally relevant signal in autobiographical text is sparse and compressible. A compact structured document captures most of it; adding more raw text adds little beyond what the structured document already delivered.

Per-subject compression comparison (5-judge primary, low-baseline slice).

The table shows baseline and every compression-related condition for each subject, with the compression ratio (source corpus tokens ÷ specification tokens, both approximate) for scale.[51]

[51] In the color-rendered PDF, low-baseline rows are tinted to mark the population of relevance; the C8 − C2a gap column is shaded to make the Spec-vs-corpus difference visible at a glance. Markdown render preserves data only.

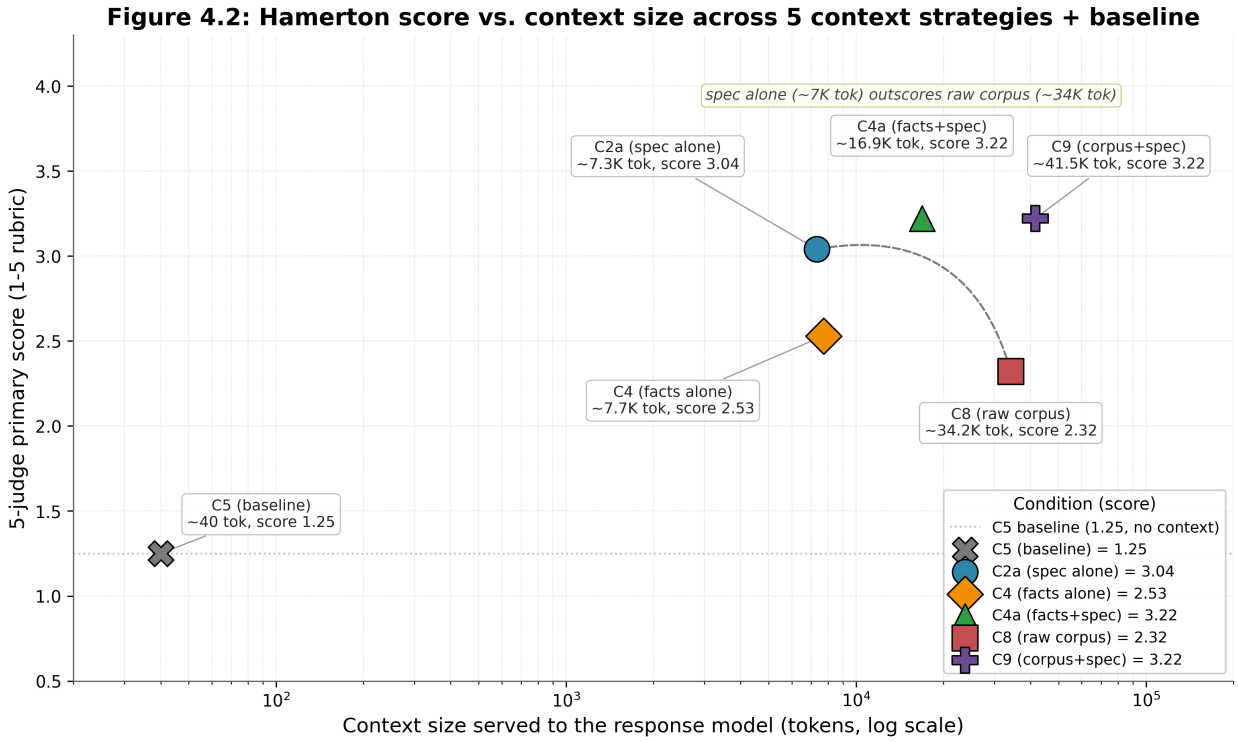


Figure 2: Figure 4.2: Score versus context size (log scale) per subject across compression-related conditions. The score climbs steeply across the first ~7K tokens of structured specification and plateaus through ~80K to 400K tokens of raw corpus, showing the compression of the behavioral signal into a small structured representation. (§4.2)

Subject	Source words (~tokens)	Compression ratio (corpus / Spec)	C5 baseline	C2a Spec (~7K tok)	C4 facts (~10K tok)	C8 raw corpus	C4a facts+Spec	C9 corpus+Spec	C8 − C2a
Hamerton	25,231 (~33K)	7×	1.26	**2.63**	2.43	2.27	2.77	**3.09**	**−0.36**
Sunity Devee	67,379 (~88K)	~13×	1.03	2.27	2.46	2.55	2.41	2.46	+0.28
Ebers	96,174 (~125K)	~17×	1.02	1.54	2.02	2.18	2.07	2.16	+0.64
Fukuzawa	139,088 (~181K)	~26×	1.67	2.35	2.67	2.74	2.78	2.78	+0.39
Bernal Díaz	187,315 (~244K)	~33×	1.70	2.27	2.41	2.55	2.48	2.53	+0.28
Bābur	422,772 (~549K)	~79×	1.76	1.91	2.03	2.05	2.01	-	+0.14
Seacole	62,467 (~81K)	~12×	1.77	2.48	2.63	2.83	2.59	2.73	+0.35
Keckley	58,742 (~76K)	~11×	1.84	2.43	2.39	2.50	2.44	2.49	+0.07
Yung Wing	66,459 (~86K)	~13×	1.88	2.22	2.13	2.42	2.40	2.50	+0.20
Mean		**~25×**	**1.52**	**2.23**	**2.35**	**2.45**	**2.44**	**2.59**	**+0.22**

All 9 low-baseline subjects show positive lift across every condition. The Spec recovers 76% of the corpus's lift at ~25× compression on average. Hamerton (smallest corpus) is the boundary case where Spec exceeds the raw corpus; Ebers (lowest baseline) is where the corpus retains its largest edge over Spec. Both are expanded in §4.2.1.

Bābur's C9 condition was excluded because the 422,772-word corpus plus the specification exceeded the response model's context window.

Once the model has the full raw corpus, adding the Spec on top contributes little at the aggregate level (~+0.09 points per-question paired); the gain is already captured at smaller context sizes by the structured representation.[52]

4.2.1 Per-question improvement rate

On the 9 low-baseline subjects, 7 of every 10 questions improve with the Spec alone (~7K tokens), within 8 percentage points of the raw corpus's improvement rate (78.3%, at ~163K tokens). The compression story holds at the per-question level: structured context produces a similar improvement rate to raw corpus context at roughly 23× less context served.

The aggregate mean score blends judge variability with response quality. A cleaner unit: out of N individual questions, how many does each condition improve over the no-context baseline? Each

[52]Per-question paired recompute on the 8-subject C9-eligible slice. The per-subject mean column at the bottom of the per-subject table reads 2.59 for C9 vs. 2.45 for C8, but the cross-subject mean of per-question paired Δs is the canonical number for the comparison.

question either improves, ties, or worsens when the condition's context is added. We report three numbers per condition: the improvement rate, the worsening rate, and the median magnitude of improvement among improved questions (with median worsening magnitude as a sanity check).

Low-baseline slice (9 subjects, 351 questions, 5-judge primary per-question means).

Condition vs. baseline	Approx. context	Improved	Tied	Worse	Improvement rate	Median Δ when improved	Median Δ when worsened
Spec only	~7K tokens	249	49	53	**70.9%**	**+1.00**	−0.40
All facts only	~10K tokens	256	44	51	72.9%	+1.00	−0.40
Raw corpus	~163K mean (33K–549K)	275	31	45	78.3%	+1.00	−0.60
All facts + Spec	~17K tokens	276	22	53	78.6%	+1.00	−0.40

When the Spec helps, the typical help is a full rubric category (+1.00 median). When it hurts, the typical hurt is less than half a category (−0.40 median). Roughly 1 in 10 questions tie; fewer than 1 in 6 worsen with the Spec alone.[53]

Multi-anchor crossings happen on 9–15% of questions when adding context to a no-context baseline; on 2–3% when the Spec is layered on top of full facts or corpus. The small mean Δ values for adding the Spec on top of facts or corpus are residues of substantial per-question movement in both directions, not uniformly small effects. The multi-anchor rate captures the categorical shifts the aggregate mean averages over (definition in §3.3.1).

Comparison	Subject set	n paired	Multi-anchor (≥2 bands)	Extreme (≥3 bands)	Mean Δ
Baseline → full pipeline (all facts + Spec)	all 14	546	13.0%	3.7%	+0.55
Baseline → all facts only	all 14	546	12.5%	4.4%	+0.47

[53] All 14 main-study subjects, matched 39-question batteries (546 questions): Spec only 58.8% improvement / 26.7% worsening; facts only 60.1% / 26.6%; raw corpus 65.2% / 23.6%; facts + Spec 65.8% / 26.4%. Pairwise question-level comparison on the low-baseline slice: raw corpus scores higher than Spec alone on 53.3% of questions (vs. 30.8% the other way, 56 ties); corpus + Spec scores higher than facts + Spec on 49.0% of questions (vs. 36.5%, 45 ties). The 7K-token facts + Spec package scores higher than the much larger corpus + Spec package on roughly one-third of questions.

Comparison	Subject set	n paired	Multi-anchor (≥2 bands)	Extreme (≥3 bands)	Mean Δ
Baseline → Spec only	all 14	546	9.0%	2.0%	+0.43
Wrong Spec → correct Spec	all 14	546	14.5%	2.4%	+0.64
Baseline → raw corpus only	13 (Bābur excl.)	507	15.4%	4.3%	+0.59
Baseline → corpus + Spec	13 (Bābur excl.)	507	14.8%	4.7%	+0.62
All facts → all facts + Spec	all 14	546	2.2%	0.9%	+0.08
Corpus → corpus + Spec	13 (Bābur excl.)	507	2.4%	0.4%	+0.03

The pattern is consistent with the §1 thesis: the Spec produces the most categorical moves where prior context is sparsest.[54]

Worked example: Hamerton Q25 across all six conditions. This question is one of the ~2.4% of corpus → corpus + Spec comparisons that produce a multi-anchor crossing, illustrating where adding the Spec on top of the full corpus produces a categorical shift. Bands follow the rubric in §3.3.

Subject: Philip Gilbert Hamerton. **Q25:** *Given Hamerton's difficulty following spoken French at Loch Awe despite years of study, what would he do about it?*

Held-out ground truth: *"This plagued me with an irritating sense of ignorance, so I looked back on my education generally, and found it unsatisfactory... I determined to acquire some substantial knowledge of modern languages, and to begin by learning French over again, so as to write and speak it easily."*

[54] All-14 figures from `docs/research/multi_anchor_rates_all_pairs_20260430.json` (script: `scripts/compute_anchor_crossing_all_pairs.py`). Row labels above use natural language; cross-reference to condition codes is in §3.2 and Appendix C. The 9-subject low-baseline slice gives somewhat higher rates (e.g., baseline → full pipeline at 18.2% on 9 subjects vs. 13.0% on all 14), reflecting that high-baseline subjects have less room to cross anchors upward at the integer-band granularity. Hamerton is a strong outlier on the corpus → corpus+Spec comparison (49% upward rate, 19 up vs. 2 down on its 39 questions) within the low-baseline slice. Per-question phenomena, including the bimodal cancellation that produces near-zero aggregate Δ on the Spec-on-info-rich pairs, are decomposed in §4.4.3 alongside memory-system layering. Multi-anchor examples: Hamerton q22 (1→3 on corpus → corpus+Spec), Hamerton q25 (1→4), Seacole q2 (2→5 on facts → facts+Spec), Yung Wing q22 (1→4 on facts → facts+Spec).

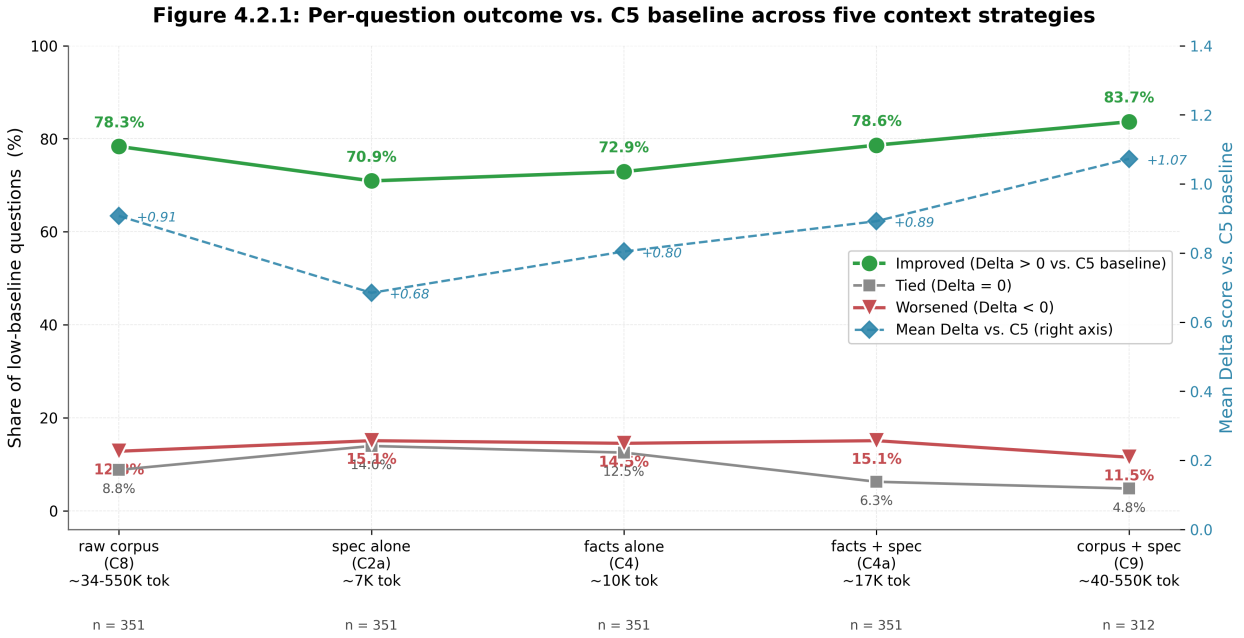


Figure 3: Figure 4.2.1: Per-question improvement rates across the five context conditions for the 9 low-baseline subjects (351 paired questions, 9 × 39). Conditions are ordered by context size: Spec alone (C2a, ~7K tokens), facts alone (C4, ~10K), facts plus Spec (C4a, ~17K), raw corpus (C8, ~163K mean), corpus plus Spec (C9, ~170K). The improved-share line spans the 70.9% to 83.7% band across conditions, with C9 highest; the tied band is intermediate; worsened stays low. Spec alone improves 70.9% of questions at roughly 23× less context than the raw corpus (78.3%); facts plus Spec matches the raw corpus's improvement rate while cutting the tied band roughly in half; corpus plus Spec produces the highest improvement rate (83.7%). Median Δ when improved is +1.00 rubric points; median Δ when worsened is −0.40 points. (§4.2.1).

Condition (natural-language)	Code	Mean	Band	Excerpt (verbatim)
No-context baseline	C5	1.0	1	“I don’t have specific information about Hamerton’s situation at Loch Awe or his particular difficulties with spoken French to give you a reliable answer.”
Spec only	C2a	3.6	3–4	“I need to flag that I don’t have direct knowledge of Hamerton or this specific incident at Loch Awe … the specification itself warns against [fabricating a prediction from the specification without grounding it in actual evidence].”

Condition (natural-language)	Code	Mean	Band	Excerpt (verbatim)
All extracted facts only	C4	1.8	1–2	"I cannot find any information in the provided facts about Hamerton's experience at Loch Awe or any difficulty he had following spoken French there. The facts do contain extensive information about Hamerton's French education. . . "
All facts + Spec	C4a	2.2	2	"I need to flag immediately that this scenario is not present in the known facts . . . I should not construct a behavioral prediction from absent evidence."

Condition (natural-language)	Code	Mean	Band	Excerpt (verbatim)
Raw corpus only	C8	1.4	1	“The autobiography provided does not contain any passage about Hamerton having difficulty following spoken French at Loch Awe or describing what he would do about it … there is no mention of Loch Awe or any specific incident.”

Condition (natural-language)	Code	Mean	Band	Excerpt (verbatim)
Corpus + Spec	C9	4.2	4	"Based on the behavioral specification and autobiography together, I can infer what Hamerton would likely do ... He would persist through disciplined, sustained practice ... The swimming episode is the clearest model: humiliated at Doncaster for falsely claiming to swim, he resolved to become genuinely expert ... He would treat the problem as one of disciplined, focused practice rather than mere exposure."

Five of six conditions refuse to engage substantively because the specific Loch Awe scenario is not narrated verbatim in the training-half corpus. C9 (corpus + Spec) is the one condition that produces a substantive prediction, integrating the Spec's axiomatic patterns (disciplined formation through shame-driven perseverance, the swimming-precedent pattern) with the autobiographical narrative to project the character pattern onto a novel scenario. The shift from refusal (band 1) to grounded prediction (band 4) on a single question is what the multi-anchor measure captures across the ~12 such questions in this comparison (2.4% of 507 paired).

Example: Hamerton, the compression story at its clearest

Hamerton is the only subject in the low-baseline slice where the Spec alone outperforms the full raw corpus. The Spec-alone lift (+1.37 over baseline) exceeds the corpus-alone lift (+1.01) on a 25,231-word source, the smallest in the study (~5× compression). Spec alone scores 2.63 vs. raw corpus at 2.27. Facts+Spec reaches 2.77. Corpus+Spec reaches 3.09, the highest compression-related score in the study, indicating Spec and corpus are complementary rather than overlapping.

When the source corpus is short enough to be sparse on its own, structured extraction adds organizational value beyond mere content. Hamerton is the boundary condition for the compression claim; the cross-subject mean still shows the corpus retaining a small edge on average.

Example: Ebers, the honest cost of compression

Ebers is the boundary case for compression's residual cost. At 96,174 words and the study's lowest baseline (1.02), every context condition lifts the score above baseline. But specification alone (1.54) underperforms raw corpus (2.18) by 0.64 points, the widest Spec-vs-corpus gap in the low-baseline slice. The 0.64 gap crosses the anchor $1 \rightarrow 2$ boundary, which by the cross-anchor rule (§3.3.1) is a real categorical shift: corpus moves Ebers's average response from refusal-or-wholly-wrong (band 1) into right-topic-wrong-prediction (band 2).

The corpus's edge is qualitative: anecdotal specificity. On Ebers Q2, the corpus cites childhood incidents (Ebers's clashes with working-class "Knoten" boys), named mentors (Langethal, Barop, Froebel), and direct autobiographical quotes. The Spec reaches the same conclusion through behavioral axioms (Truth Over Embellishment, the reasoning signature), principled reconstruction rather than lived evidence:

> **C2a (Spec only):** "Ebers would show receptiveness to this critique *if* it comes from a mentor he has already internalized as exemplary ... He would accept that his lived experience is necessarily limited in scope, but resist the implication that his *method* of understanding is fundamentally flawed."
>
> **C8 (raw corpus):** "Ebers explicitly acknowledges gaps in his understanding ... He reflects on his childhood fights with the 'Knoten' (working-class boys) and later recognizes the injustice of class-based mockery: 'If they had called us boobies we should probably have called them blockheads ... Children don't fight regularly with those whom they despise.' Ebers demonstrates profound deference to figures like Langethal, Barop, and Froebel."

The gap is the rubric registering anecdotal texture over principled reconstruction. Compression captures the bulk of the signal; the residual on Ebers is autobiographical specificity the Spec abstracts into axioms by design.

Why this matters for deployment.

At any scale where a per-user full autobiography cannot be served into context on every query, the compression result is what makes personalization operationally tractable. The specification's

7K-token footprint is within normal per-request context budgets. A 100,000-to-400,000-word corpus is not. **The specification recovers 76% of the corpus's predictive benefit at 4% of the context cost; the remaining ~24% is achievable but only at production-prohibitive context sizes.**[55]

4.2.2 Three statistical signatures

The Spec produces three distinct signatures depending on what context the model already has. On an empty baseline, it re-ranks: different questions become the well-answered ones ($\rho = 0.27$). On top of facts or corpus, it lifts most answers similarly without re-ranking ($\rho = 0.71$–0.72). On top of the Spec alone, adding facts produces a mix: some re-ranking, some uniform lift ($\rho = 0.62$).

The pre-vs-post Spearman ρ across questions captures how much the Spec changes which questions get answered well. A low ρ means the question ranking shifted substantially (re-ranking). A high ρ means ranking was preserved and most answers were lifted by a similar amount (uniform lift).

Condition pair	ρ	What this means in plain terms
Baseline → all facts + Spec	0.27	**Re-ranking.** Different questions become the well-answered ones; the Spec changes which questions the model can handle.
All facts only → all facts + Spec	0.72	**Near-uniform lift.** The same questions stay strong; the Spec lifts most answers by a similar amount.
Raw corpus only → corpus + Spec	0.71	**Near-uniform lift.** Same picture as facts: ranking preserved; the Spec adds a roughly even lift across questions.
Spec only → all facts + Spec	0.62	**Partial re-ranking.** Mixed: adding facts on top of the Spec reorders some questions and uniformly lifts others.

The re-ranking signature on the empty-baseline comparison could mean either that the Spec lets

[55] Raw per-subject data at `results/global_<subject>/c8_c9_results.json` and `results/global_<subject>/results_v2.json`. The compression analysis is in `scripts/recompute_5judge_primary.py`; per-question improvement rates are computed at render-time in `scripts/generate_fig_4_2_1.py`. Figure 4.2 plots score versus context size (log scale) per subject and shows the steep initial climb and long plateau.

the model answer a different set of questions, or simply that baseline scores cluster near the rubric floor where reordering is structurally easier; a future test with a non-floor-anchored baseline would distinguish the two readings.

4.3 Mechanism: Correct Content, Not Format

Hypothesis tested in this section (H3 from §1.2): The benefit comes from the content of the correct specification for the correct person, not from the mere presence of a structured prompt. A random other person's specification, applied in its place, does not reproduce the effect.

Wrong Spec drops accuracy below the no-context baseline (adversarial v1, $\Delta = -0.25$) and 0.60 points below correct Spec ($\Delta = +0.35$), more than half a rubric-anchor category. Random-derangement v2 sits between (+0.15). If structure alone drove the effect, mismatched specs would recover most of the lift. They recover at most part of it (random v2 = +0.15) or actively degrade below baseline (adversarial v1 = −0.25). The +0.15 random lift is coincidental content overlap, not a structure-alone effect: occasionally the wrong Spec's pattern happens to predict the same surface behavior as the correct subject's by chance (Example B below traces one such case; the scoring rubric's coverage of right-action-wrong-logic situations is discussed in §6.2). The asymmetry also speaks to the specifications themselves: they are sufficiently person-specific that mismatches register as content errors, not as inert structural prompts.

On the 13 global subjects with complete 5-judge primary coverage, three conditions test whether content matters:

Condition	Mean Δ vs. C5 (5-judge primary, 13 globals)	Reading
C2a (correct Spec)	**+0.35**	matched content increases representational accuracy
C2c v2 (random derangement, seed-fixed)	**+0.15**	partial improvement; dominated by floor effects on low-baseline subjects
C2c v1 (fixed derangement, cultural/temporal distance maximized)	**−0.25**	adversarial mismatch degrades representational accuracy below the no-context baseline

The two wrong-Spec variants differ by construction. **v1 (fixed derangement)** is a hardcoded pairing designed so each subject receives the specification of a culturally- and temporally-distant other (for example, Ebers the 19th-century German Egyptologist receives Equiano the 18th-century West-African/British autobiographer; Seacole the 19th-century Jamaican nurse receives Bernal Díaz the 16th-century Spanish conquistador).[56] **v2 (random derangement)** is a seed-fixed random permutation in which no subject receives its own specification but pairings can land culturally-close; this tempers the aggregate drop. Reporting both shows that even a random wrong-Spec barely

[56] Pairing logic in `scripts/run_global_rerun.py`.

improves on no context, and an adversarial wrong-Spec actively hurts.[57]

Three mechanism types.

Three distinct mechanisms produce the correct-specification improvement across the study data. Each has a characteristic wrong-specification failure mode, illustrated in the matched examples below.

1. **Identity disambiguation.** When the baseline model cannot determine which person is being asked about, the specification provides enough content (temporal markers, cultural domain, documented life events) to resolve the identity and anchor the reasoning frame.
   - *Wrong-Spec failure mode:* the model either detects the mismatch explicitly and refuses to predict, or anchors on the wrong person's pattern and produces a coherent but off-target prediction.
2. **Directional correction.** When retrieved facts suggest a generic-default prediction that contradicts the subject's actual pattern, the specification overrides the generic with the subject-specific.
   - *Wrong-Spec failure mode:* the model applies the wrong person's pattern; depending on how close that pattern happens to be to the target subject's, the prediction is either directionally wrong in a new way or coincidentally correct (the wrong person's pattern happens to predict the same surface behavior on this particular question, for different underlying reasons; Example B below is one such case).
3. **Interpretive inference.** When retrieved facts do not include direct evidence for the specific question, the specification provides interpretive scaffolding to generalize from established character patterns to the new situation.
   - *Wrong-Spec failure mode:* the model detects the mismatch and refuses, or applies wrong-person scaffolding and produces a low-quality prediction.

Response-level evidence: when the model engages with the Spec, and when it does not.

Three signals from response text confirm that content matters more than structure.

Spec-tag citation gap. Models cite Spec-specific tags (anchor IDs, axiom references, predictive-template labels) on **78.6%** of correct-Spec responses but only **50.0%** of wrong-Spec responses.[58] The 28.6-point gap is a lower bound on the content effect; models may draw on Spec content without literally quoting tag IDs.

Models can detect when a specification does not fit the named subject. Across 587 wrong-Spec responses, **60.6% explicitly flagged the content mismatch** (example: *"This is a behavioral model of a 16th-century Central Asian military ruler, almost certainly Bābur"*). 36.5% attempted to apply the mismatched content and produced low-quality predictions; 3% hedged or were ambiguous.[59] The detection signal is interpretive content (temporal markers, cultural domain,

[57] Both wrong-Spec deltas hide per-subject variation. The full per-subject table, the count of subjects with positive vs. negative outcome under each protocol, and the adversarial-vs-random pairing sensitivity are in §4.6.5.

[58] Data at `docs/research/spec_activation_analysis.json`.

[59] 587 wrong-Spec responses: 507 from the v2 random-derangement protocol on the 13 global subjects, plus 80 from the v1 adversarial protocol on Hamerton across all five battery tiers. Validated against a 30-response stratified manual spot check. Full breakdown: 60.6% explicit mismatch flag, 36.5% applied mismatched content, 2.0% hedged implicitly, 0.9% ambiguous.

documented life events) being inconsistent with what the model already knows about the named subject. Specifications are anonymized (§3.7), so the model has no surface name cue to compare against; only interpretive content.

Hedging persists under wrong-Spec. Correct-Spec conditions eliminate baseline hedging: 41.2% → 0.4% under the broader-pattern classifier.[60] On wrong-Spec, refusal patterns persist at the 60.6% explicit-detection rate above. The hedging-elimination is content-specific, not structure-specific.

Per-question matched comparison.

The three §4.1 examples extend directly into mechanism-by-mechanism wrong-Spec contrasts. Same subject, same question, same response model; the only change is the specification content. Per-question swings are larger than the +0.60 mean gap in both directions. Correct-Spec C4a responses are in §4.1 for reference.

Example A (wrong-Spec), Ebers Q7: identity disambiguation fails

Context. Clean-mismatch case: a Spec that describes someone other than the named subject, served under the v1 fixed-derangement protocol. The model detects the mismatch and declines: 1.60 vs. correct Spec 3.60, a 2.00-point drop.

Subject + question. Same as §4.1 Example A (Ebers, self-sacrifice and educational institutional success). Battery-question target name: Ebers (explicit in the prompt).

Wrong Spec served. Equiano's specification, anonymized.[61] The response cites two of Equiano's anchors: A4 Systemic Indictment and A9 Economic Self-Determination.

Wrong-Spec response.

> *"I need to be direct: I don't have reliable access to specific statements by Ebers on this question. The behavioral specification you've provided is detailed and rich, but it doesn't contain material where this person directly addresses the relationship between self-sacrifice and educational institutional success. ... Given A4 (Systemic Indictment) — that institutionalized greed is the generative engine of cruelty ... a teacher sacrificing their own economic security, health, or family time does not fix a system structured by avarice."*

Reading. The model detected the mismatch between the named target in the question (Ebers, a 19th-century German Egyptologist) and the interpretive content of the anonymized specification (anti-slavery and economic-freedom anchors, which are Equiano's). It named the served anchors correctly, reasoned from Equiano's framework, and declined to produce a prediction about Ebers. The identity-disambiguation mechanism that enabled the correct Spec's lift in §4.1 Example A did not fire because the Spec content is not about Ebers. The response falls near the rubric floor because no prediction was made.

[60] Strict-pattern classifier yields 28.8% → 0.0%. Rule definitions in §1.3 footnote.
[61] Anonymization protocol in §3.7. Pairing logic in `scripts/run_global_rerun.py`.

Example B (wrong-Spec), Bernal Díaz Q16: content convergence across genuinely different frameworks

Context. Coincidence case: random pairing happens to converge on the same surface prediction by different logic. v1 fixed-derangement protocol. Wrong Spec scored 4.60 vs. correct Spec 4.80, a 0.20-point near-tie. This is what produces the wrong-Spec aggregate's drift above zero on the random-pairing v2 protocol.

Subject + question. Same as §4.1 Example B (Bernal Díaz, Cortes and offered physical assistance). Battery-question target name: Bernal Díaz.

Wrong Spec served. Sunity Devee's specification, anonymized.[62] The response cites three of her anchors: A5 Relational Identity, A9 Simplicity as Virtue, P5 Virtue through Embodied Practice.

Wrong-Spec response.

> *"Based on this behavioral specification, the author would likely decline the offered assistance ... Accepting help during a strenuous activity would read as: a failure of self-discipline, an implicit claim to comfort or ease that contradicts spiritual seriousness, a reliance on others that undermines the embodied practice of virtue (A9, P5) ... Physical exertion endured without assistance becomes evidence of character."*

Reading, not parroting. The two specs are genuinely different frameworks. Bernal Díaz's actual anchors include A1 Divine Mandate, A2 Civilizational Hierarchy, A4 Loyalty Architecture, A5 Forward Compulsion, A6 Gold as Proof, and A10 Charismatic Override: a conquistador's martial-providential register. Sunity Devee's anchors cited here (A5 Relational Identity, A9 Simplicity as Virtue, P5 Virtue through Embodied Practice) are an ascetic-devotional register. Direct anchor-to-anchor comparison across the two specs finds zero substantive mirroring. On the specific question of refusing offered physical help, the two frameworks converge by different logics. The correct Spec (Bernal Díaz) predicts refusal because accepting help would signal weakness to followers and violate performative self-reliance (A4 + A5 in the conquistador register). The wrong Spec (Sunity Devee) predicts refusal because accepting help would compromise physical discipline and violate simplicity-as-virtue (A9 + P5 in the devotional register). Different moral architectures, same overt behavior.

Why the correct Spec still scored higher, 4.80 vs. 4.60. Both conditions predicted the right surface action. The 0.20-point gap is judge preference for rationale specificity that matches the ground-truth passage's tone. The correct Spec's "symbolic or morale-signaling purpose" rationale maps onto a battlefield memoir's register. The wrong Spec's "spiritual seriousness, devotional simplicity" rationale predicts the same action but in a register alien to Cortes on the steps of the Templo Mayor. Judges reward tonally-aligned rationale. The convergence is real but costs precision.

Why this matters for the aggregate numbers. This is the coincidence case that lifts the wrong-Spec aggregate above zero. When pairing happens to land on subjects whose behavioral patterns converge on the same surface prediction for a given question, wrong-Spec lift is real on that question, not an artifact. Across the 507-response v2

[62] Anonymization protocol in §3.7. Pairing logic in `scripts/run_global_rerun.py`.

cohort (13 globals × 39 questions), mismatch-loss dominates on the adversarial v1 pairing (aggregate Δ −0.25) and roughly balances on the random v2 pairing (aggregate Δ +0.15). Example B is one of the roughly 5-10% of questions where content coincidence produces correct-surface, wrong-logic predictions.

Example C (wrong-Spec), Seacole Q2: interpretive inference fails

Context. Clean-mismatch case: the served Spec is from a different century and culture than the named subject. v1 fixed-derangement protocol. Wrong Spec scored 1.40 vs. correct Spec 5.00, a 3.60-point drop.

Subject + question. Same as §4.1 Example C (Seacole, delirious patient). Battery-question target name: Mary Seacole.

Wrong Spec served. Bernal Díaz's specification, anonymized.[63] The response identifies the wrong Spec by its interpretive content alone and names four of his anchors: A1 Divine Mandate, A2 Civilizational Hierarchy, A4 Loyalty Architecture, A5 Forward Compulsion.

Wrong-Spec response.

> *"The specification you've provided describes someone from the 16th-century Spanish conquest period — the epistemic anchors, loyalty architecture, divine mandate framing, and forward compulsion all point to that historical context. Mary Seacole (1805-1881) was a Creole nurse and travel writer from Jamaica who served in the Crimean War. These are entirely different people from different centuries with different frameworks. . . . I don't have documented instances of this. I could speculate based on her memoir's tone and her known character, but that would be inference, not behavioral specification."*

Reading. The model detected the mismatch between the named target in the question (Mary Seacole, a 19th-century Jamaican Creole nurse) and the anonymized content of the served specification (16th-century Spanish conquest anchors). It named the anchors it was seeing, placed Seacole in a different era and role, and refused to apply the mismatched content. The interpretive-inference mechanism that produced §4.1 Example C's correct-Spec 5.00 score does not fire: without Seacole's actual character pattern in context, the model would not generalize from an unrelated conquistador's framework to her delirious-patient scenario.

Summary of the three examples.

[63] Anonymization protocol in §3.7. Pairing logic in `scripts/run_global_rerun.py`.

Example	Mechanism (correct Spec)	C4a (correct)	C2c v1 (wrong)	Drop	Wrong-Spec pattern
A (Ebers Q7)	Identity disambiguation + interpretive inference	3.60	1.60	**−2.00**	Explicit mismatch flag; declined prediction
B (Bernal Díaz Q16)	Directional correction	4.80	4.60	**−0.20**	Coincidental content overlap; wrong-Spec prediction matches
C (Seacole Q2)	Interpretive inference	5.00	1.40	**−3.60**	Explicit mismatch flag; declined prediction

Two of three examples show large drops (−2.00 to −3.60 points) when the content does not fit. The third shows near-zero drop, but only because the wrong Spec's content happens to predict the same surface behavior. That asymmetry, clean mismatches versus coincidental overlaps, is exactly what the aggregate Δ numbers reflect: the adversarial-pairing v1 aggregates to −0.25 because most questions are mismatch cases, and the random-pairing v2 aggregates to +0.15 because random pairings more often hit content-proximity combinations like Example B.[64]

4.4 Memory-system composition

Hypothesis tested in this section (H4 from §1.2): The Behavioral Specification layers cleanly on top of memory-system retrieval rather than replacing it. The Spec contributes representational accuracy beyond what retrieval alone provides, and that contribution decomposes into per-question patterns characteristic of each retrieval architecture.

4.4.1 Cross-system retrieval: providers do not converge

On 35.9% of instances, any two memory systems share zero facts in their top-10. They retrieve no overlapping facts at all on the same question. Averaged across all ten system pairings, the mean overlap is 8.3%. Recall benchmarks like LongMemEval and LOCOMO measure whether a system can retrieve a previously-stored fact, and the four commercial systems we tested score within a few percentage points of each other on those benchmarks. Representational

[64] Raw per-judge data and full response text at `results/global_<subject>/results_v2.json` (wrong-Spec responses) and `results/global_<subject>/judgments_v2.json` (per-judge scores). Analysis scripts at `scripts/compute_wrong_spec_5judge.py` and `scripts/compute_wrong_spec_per_subject.py`.

accuracy and behavioral prediction operate at a different layer, where the relevant question is which facts matter for a specific interpretive task.

Setup. Four commercial memory systems (Mem0, Letta[65], Supermemory, Zep) and Base Layer's own zero-cost retrieval substrate (MiniLM-L6-v2 + ChromaDB) were tested under two configurations. We ran both to separate two layers of variation: how each system *ranks* facts from a fixed pool (controlled), and whether each system's *ingestion* pipeline adds further divergence on top (native).

- **Controlled configuration.** Each system is given an identical pre-extracted fact pool drawn from the training half of each subject's corpus. Holds the input constant across systems; differences trace to retrieval and presentation policy alone.
- **Native configuration.** Each system ingests the raw training corpus through its own production pipeline, as in deployment. Measures the full end-to-end system.

Convergence test. No, the provider layer does not converge on relevance. On 35.9% of (system pair, question) instances two systems share zero facts in their top-10s; on 65.6% they share one or fewer; the mean pairwise overlap across the ten system pairs is 8.3%.[66] Convergence on top-K under identical input would have been evidence of a shared interpretive substrate. The lack of convergence suggests the rankings reflect provider-specific design choices rather than a shared theory of which facts the question is asking for. The controlled configuration isolates the ranking layer as a load-bearing source of representational accuracy: which facts the system surfaces determines, before any reading model engages, what the response can be about.

For any two systems on the same question, the fraction of retrieved facts that appear in both lists is called the **Jaccard similarity**.[67] A Jaccard of 1.0 means identical top-10s; 0.0 means no shared facts. We compute it for each of the ten system pairs (Base Layer, Letta, Mem0, Supermemory, Zep) on each of 546 questions, in the controlled configuration where every system reads the same all-facts pool.

Pairwise Jaccard similarity (controlled configuration, n=546 questions per pair):

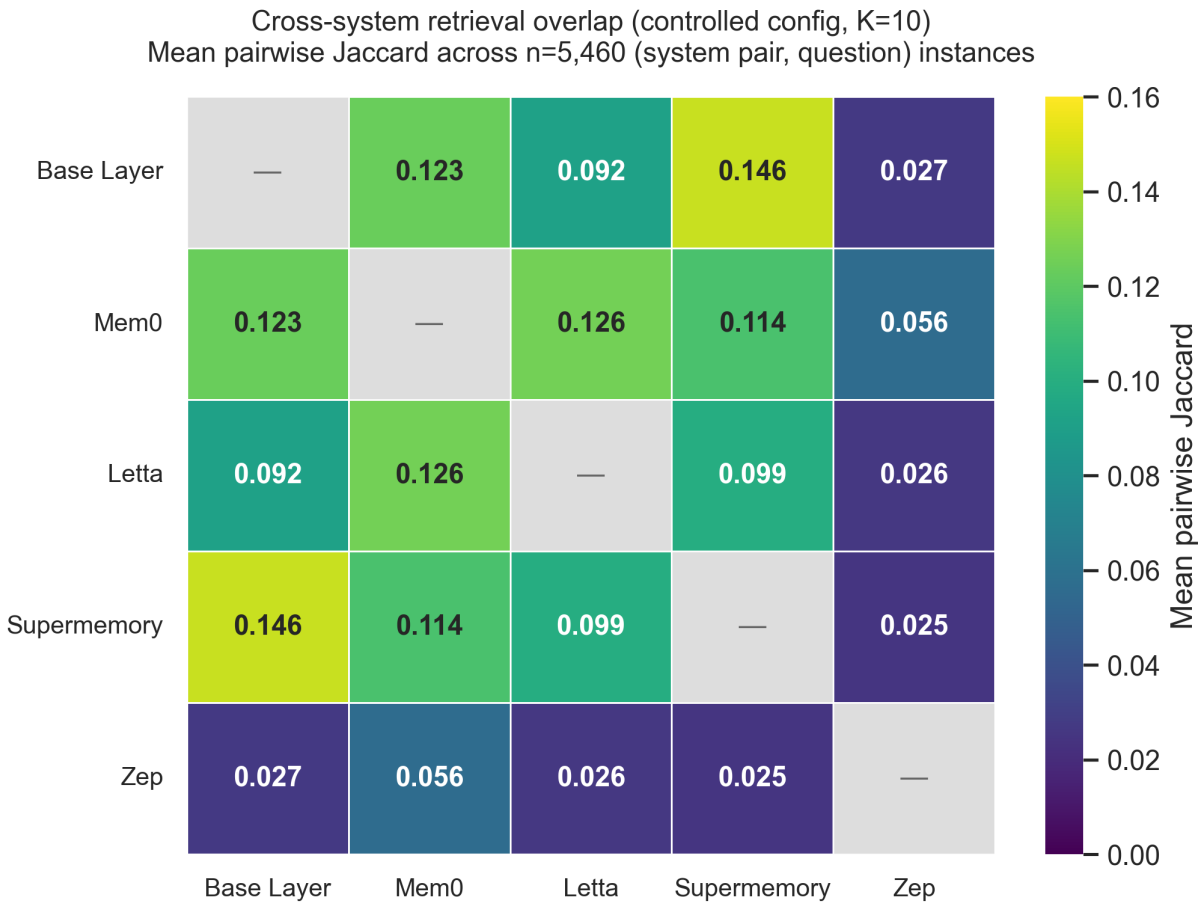
System pair	Mean Jaccard
Base Layer ↔ Supermemory	0.146
Mem0 ↔ Letta	0.126
Base Layer ↔ Mem0	0.123
Mem0 ↔ Supermemory	0.114
Letta ↔ Supermemory	0.099
Base Layer ↔ Letta	0.092
Mem0 ↔ Zep	0.056
Base Layer ↔ Zep	0.027
Letta ↔ Zep	0.026

[65] Letta's archival-retrieval path is reported here. Letta also exposes a second path (a persistent memory block its agent self-edits during ingestion, the MemGPT design), which is architecturally unlike a retrieval path and is reported separately in §4.5.

[66] Share-zero / share-≤1 fractions computed across all 14 main-study subjects × 39 behavioral-prediction questions × 10 system pairs = 5,460 (system pair, question) instances under the controlled retrieval configuration. Excluding Hamerton (13 globals × 39 × 10 = 5,070 instances) shifts share-zero to 36.7% and share-≤1 to 66.6%. Restricting to the four commercial systems (six pairs, BaseLayer excluded) shifts share-zero to 40.4% (14 subjects) or 41.0% (13 globals); every cut shows substantial top-K divergence on identical input. Reproducibility script at `scripts/analyze_retrieval_overlap.py`; data at `docs/research/retrieval_overlap_analysis_20260501.json`.

[67] Jaccard similarity is computed as the size of the intersection divided by the size of the union of the two top-K sets.

System pair	Mean Jaccard
Supermemory ↔ Zep	0.025
Mean across pairs	**0.083**



Mean Jaccard 0.083 across 10 system pairs; lowest pair Supermemory–Zep 0.025; highest Base Layer–Supermemory 0.146.
Source: scripts/analyze_retrieval_overlap.py.

Figure 4: Figure 4.4.1: Cross-system retrieval overlap. Mean pairwise Jaccard between every pair of memory systems on the controlled retrieval configuration (n=5,460 = all 14 main-study subjects × 39 behavioral-prediction questions × 10 system pairs). The diagonal is grayed; cells below the diagonal mirror cells above. Highest pair Base Layer–Supermemory at 0.146; lowest Supermemory–Zep at 0.025. Zep's row is uniformly low (graph-traversal scoring overlaps weakly with embedding-similarity retrieval). Mean across pairs 0.083.

Pairs involving Zep are the lowest. Zep ranks facts by traversing relationships in a knowledge graph; the other systems rank by how close a fact's meaning is to the question (embedding similarity). The two ranking approaches surface different facts on identical input, so Zep overlaps weakly with everyone else.

A note on counting. We count *unique facts* in each top-10: if the same fact text appears more than

once in a system's returned list, we count it once. Some systems return duplicates because the underlying scoring traverses graph relationships and surfaces the same fact under different paths. Zep's controlled retrieval returns 10 entries per question, but only about 9.8 are unique on average. Letta's returns 10 entries but only about 3.5 are unique; the same fact appears multiple times under different graph paths. Counting only unique facts, Letta's effective retrieval depth is shallower than its raw count suggests.

Comparing each commercial system to the average of the others' rankings: Mem0 is the closest match (mean Jaccard 0.105 across Mem0's four cross-pairs), followed by Supermemory (0.096), Letta (0.086), and Zep (0.034). Mem0 retrieves what the average of the other systems retrieves; Zep retrieves a different set.[68]

Under the native pipeline, the lack of shared facts is even more pronounced. Each system returns its retrieval in a different format: Mem0 returns third-person summary sentences, Letta returns raw multi-sentence passages, Supermemory returns atomic facts, Zep returns rows extracted from a knowledge graph. Because the four systems return content in different shapes, any two systems share zero exactly-matching facts on the same question, and pairwise overlap drops to 0.000 across all four native pairs. A semantic-similarity check that lets two facts count as a match when they share roughly the same content (rather than the same exact wording) raises the native overlap only marginally, to 0.004 at the near-duplicate threshold (cosine ≥ 0.85) and 0.016 at the loose topical threshold (cosine ≥ 0.70). The divergence is structural, not a surface-form artifact. The same check applied to the controlled configuration above also leaves the divergence intact; full sensitivity grid in §4.6.6.

Whether divergent facts produce divergent answers is the practical test of whether retrieval differences matter. §4.4.2 (aggregate Δ across systems and ingestion paths) and §4.4.3 (per-question patterns where the layer's effect is concentrated) take that up.

4.4.2 Layering the Spec: aggregate Δ across systems and ingestion paths

Layered on top of three of four commercial memory systems (Mem0, Letta, Zep), the Behavioral Specification produces a net-positive aggregate Δ across the 14 main-study subjects. Wilcoxon signed-rank confirms direction at $\alpha = 0.01$ on four (system, configuration) cells. The aggregate Δ on every system is the balance of per-question patterns; that decomposition is in §4.4.3 (where Supermemory's near-zero aggregate Δ is also unpacked).

Conditions compared. Within each system in each configuration: - **C1 (retrieval only):** the memory system's retrieval served as context; no Behavioral Specification. - **C3 (retrieval + Spec):** the same retrieval plus the full Behavioral Specification.

The Spec-effect for that system is **Δ_spec = mean(C3 retrieval + Spec) − mean(C1 retrieval only)**, aggregated per subject, then averaged across subjects. If the specification helps

[68] Mean pairwise Jaccard across all ten pairs is 0.083 raw, 0.088 after lowercase + whitespace normalization. Per-subject Jaccard varies from 0.043 (Equiano) to 0.115 (Hamerton); per-category variation is small (0.076–0.093 across decisions, values, relationships, conflict, learning). The divergence is a property of provider ranking, not of question type or subject. Full per-pair, per-subject, per-category breakdowns at `docs/research/retrieval_overlap_analysis_20260501.json`. Reproducibility script at `scripts/analyze_retrieval_overlap.py`.

memory-system performance, Δ_spec is positive across systems.

Aggregate and per-question results (5-judge primary, all 14 main-study subjects).

System	Config	Δ_spec	% subjects improved	% questions up ≥1 anchor	% questions up ≥2 anchors
Mem0	controlled	+0.12	71%	24.0%	3.1%
Mem0	native	+0.33	71%	**37.1%**	**8.1%**
Letta (archival)	controlled	+0.20	86%	28.1%	6.1%
Letta (archival)	native	−0.02	36%	20.5%	0.9%
Zep	controlled	+0.19	93%	29.3%	5.3%
Zep	native	+0.33	93%	**35.3%**	**7.5%**
Supermemory	controlled	−0.05	36%	18.6%	2.5%
Supermemory	native[69]	−0.01	43%	19.0%	1.4%
Base Layer substrate	controlled	+0.08	64%	26.7%	3.9%

Headline numbers report the all-14 panel. Three of four commercial systems produce a positive aggregate Δ_spec under at least one configuration; Supermemory aggregates near zero under both. Base Layer's substrate produces the smallest positive Δ; the interpretive improvement comes from the specification itself, not from retrieval choices.

Aggregate Δ_spec masks substantial per-question variance. Every system lifts 18–37% of questions by at least one rubric anchor, and 1–8% by two or more anchors. Both are categorical changes per the §3.3.1 cross-anchor interpretation rule. Mean Δ averages over a population where the Spec produces categorical change on a subset of questions and small adjustments or losses elsewhere. Even Supermemory, with a near-zero aggregate, lifts 18.6% of questions by at least one anchor (controlled). The per-question redistribution is decomposed in §4.4.3.

Wilcoxon signed-rank confirms direction at $\alpha = 0.01$ for Zep controlled ($p = 0.0004$), Letta controlled ($p = 0.0017$), Mem0 native ($p = 0.0088$), and Zep native ($p = 0.0015$). Mem0 controlled is significant at $\alpha = 0.05$ ($p = 0.017$). Letta native, Supermemory (both configurations), and Base Layer substrate are not significant at $\alpha = 0.05$. The 9-subject low-baseline slice was computed but is statistically underpowered at the effect sizes these systems show, so we do not lead with it; per-subject low-baseline detail is in the footnote.[70]

[69] Supermemory native covers 10 of 14 subjects after paid-tier rerun failures on Bābur, Bernal Díaz, Cellini, and Rousseau (n=221 paired questions vs. ~546 expected). Per-question rates conditional on the 10 covered subjects.

[70] Wilcoxon signed-rank on C1 vs. C3, all-14 panel: Zep controlled $p = 0.0004$, Letta controlled $p = 0.0017$, Mem0 native $p = 0.0088$, Zep native $p = 0.0015$ (all four robust at $\alpha = 0.01$). Mem0 controlled $p = 0.0166$ (significant at $\alpha = 0.05$, not at $\alpha = 0.01$). Letta native, Supermemory (both configurations), and Base Layer substrate are not significant at $\alpha = 0.05$ on either the all-14 or low-baseline-9 slice. Low-baseline 9-subject slice, controlled configuration: Mem0 +0.10 (6/9 improved), Letta +0.17 (8/9), Zep +0.17 (9/9), Supermemory −0.01 (4/9), Base Layer +0.08 (6/9). Native: Mem0 +0.32 (7/9), Letta −0.04 (4/9), Zep +0.30 (9/9), Supermemory −0.03 (4/9). The Supermemory native aggregate covers all 14 subjects under a paid-tier rerun completed 2026-04-23; 30 provider-failure placeholders (Augustine 2 q, Equiano 28 q) are scored at the rubric floor and treated as scored data, not missing data; qualitative story holds either way. Per-system per-subject per-judge scores at `results/global_<subject>/*_judgments*.json`.

Native ingestion shapes how much room the Spec has to contribute on top, and that interaction varies by system. When the input is held constant (controlled), four of five systems produce a positive Δ_spec; the specification's contribution is visible on top of an identical fact pool. When each system ingests its own way (native), the systems split. Mem0 and Zep increase under native (Mem0 $+0.12 \rightarrow +0.33$, Zep $+0.19 \rightarrow +0.33$). Letta decreases sharply ($+0.20 \rightarrow -0.02$). Supermemory stays roughly flat ($-0.05 \rightarrow -0.01$). Mem0's controlled-to-native lift of $+0.21$ is the largest among the four commercial systems, with Zep's at $+0.14$. The mechanism for these splits is in §4.4.3: the Spec helps retrieval-based systems on interpretation-heavy questions they were not designed for, sometimes hurts on literal-recall questions retrieval already answered, and induces principled refusals on questions where retrieved facts cannot ground a prediction.

4.4.3 Where the Spec helps, where it hurts, and which question types route to each

Three patterns of Spec-retrieval interaction emerge across all five systems tested. The same three patterns produce positive aggregate Δ_spec on three commercial systems and near-zero on Supermemory. What changes between systems is how the Spec helps or hurts across the question battery, not the patterns themselves.

The three patterns:

1. **Interpretive supply.** When retrieval underdetermines the answer, the specification provides interpretive scaffolding to generalize from established character patterns to the specific question. *Increases representational accuracy on the question.*
2. **Over-theorization.** When retrieval already supplies the plain answer, the specification can pull the response toward interpretive depth that the question does not call for. *Decreases representational accuracy on the question.*
3. **Spec-induced refusal.** Specification axioms (which vary by subject; in this study, dignity, honoring-testimony, and epistemic-integrity axioms across different subjects) can trigger a meta-refusal where the model declines to predict. The current content-match rubric cannot distinguish principled refusal from a wrong prediction (§3.3.6). *Lowers the measured rubric score; whether it lowers actual representational accuracy depends on whether refusal was the correct behavior on that question.*

How each system shows these patterns:

- **Mem0:** Mem0's atomic-fact retrieval surfaces clean, tightly-scoped facts. When the question requires interpretation beyond those facts, Pattern 1 lifts the response by supplying the framework. When the question is literal-recall and the atomic facts already resolve it, Pattern 2 takes over: the Spec abstracts away from the plain answer Mem0 already provided. Pattern 3 occurs but registers as a smaller rubric penalty because Mem0's retrieval often hedges near the rubric floor on refusal-triggering questions already, leaving little room for the Spec's refusal axioms to drag the score further down.
- **Letta archival:** Letta's controlled retrieval returns 10 entries per question but only about 3.5 are unique (graph traversal repeats facts). When those few unique facts align with the

Per-system anchor-crossing analysis at `docs/research/per_system_anchor_crossing_20260427.md` and `.json`. A hedging-reduction pattern surfaced during the memory-system analysis but did not track the Spec-effect magnitude cleanly across systems; the content-specific reading from §4.3 holds.

Spec's interpretive frame, Pattern 1 produces large lifts. When they do not align, the Spec either over-theorizes (Pattern 2) on the available content or refuses (Pattern 3) for lack of grounding evidence.

- **Zep:** Zep's temporal-graph retrieval returns relational structure rather than atomic facts. The relationship-rich context tends to suit the Spec's interpretive framing across question types. Zep shows the most balanced pattern distribution: Pattern 1 lifts roughly as often as Pattern 2 hurts, with fewer large regressions than the other commercial systems. Pattern 3 surfaces on questions where Zep's C1 retrieval is productive enough that the Spec's refusal axioms convert it into a measurable drop (the Ebers Q18 reproduction below is one such case).
- **Supermemory:** Supermemory's strong embedding retrieval gives the highest C1 mean across systems (≈ 2.61 controlled), so it more often supplies a plain answer to the model on its own. This shifts the balance toward Pattern 2 (over-theorization on questions retrieval already answered) and Pattern 3 (the Spec's evidentiary axioms convert a productive C1 answer into abstention). Supermemory's near-zero aggregate Δ_spec is the visible signature of helps and hurts roughly canceling.
- **Base Layer:** Base Layer's MiniLM + ChromaDB substrate is the leanest retrieval tested. The Spec carries proportionally more of the interpretive load: Pattern 1 dominates on interpretation-heavy questions. Pattern 3 surfaces on questions where the lean retrieval cannot ground a prediction at all.

Per-system per-subject paired-delta distributions and the full per-system breakdown are in Appendix B.11.[71]

The examples below draw from Supermemory because its near-zero aggregate Δ_spec makes the helps and hurts most legible at the per-question level: 110 of 546 paired questions cross by $|\Delta| \geq 1.0$ on the 5-point rubric, splitting 57 helps (mean swing $+1.55$) and 53 hurts (mean swing -1.38).[72] Each anchor example is paired with a same-mechanism case from a different commercial system to confirm reproduction.

Pattern 1: Interpretive supply.

Anchor example: Fukuzawa Q26 (Supermemory)

Scores: retrieval only 2.00 → retrieval + Spec 4.20 (Δ +2.20).

Question: What is Fukuzawa's attitude toward visiting friends whose households have questionable reputations? **Held-out ground truth:** *"So I feel no hesitation in paying a visit where there is a young daughter in the house or where the young wife is staying by herself… I am not put out by the gayety."*

[71] Per-system per-subject paired-delta table (Mem0, Letta archival, Zep, Base Layer × representative subjects) in Appendix B.11. Every row is a mixture of increases and decreases at the per-question level: Mem0 Yung Wing aggregate Δ +0.33 = 21 increases / 10 decreases; Mem0 Keckley aggregate Δ −0.02 = 12 increases / 13 decreases; Letta archival Hamerton +0.42 = 19 / 7; Zep Seacole +0.47 = 20 / 7 (and 0 large regressions); Base Layer Yung Wing +0.29 = 19 / 7. Even strong-positive aggregates contain regressions; even near-zero aggregates resolve into substantial counts in both directions. Full data and recompute script at `docs/research/per_system_anchor_crossing_20260427.json` and `scripts/_table_4_6_5judge_recompute.py`.

[72] Across 546 paired main-study questions with 5-judge primary coverage on both C1 and C3: 110 questions (20.1%) have $|\Delta(\text{C3} - \text{C1})| \geq 1.0$ on the 5-point rubric, splitting 57 helps ($\Delta \geq +1.0$, mean $+1.55$) versus 53 hurts ($\Delta \leq -1.0$, mean -1.38). Scaffold values: `4_4_2_supermemory_helps_n` (57), `4_4_2_supermemory_hurts_n` (53), `4_4_2_supermemory_paired_total_n` (546).

C1 (Supermemory retrieval alone), mean 2.00: *"The information does not directly address Fukuzawa's attitude toward visiting friends with questionable household reputations... a selective and cautious approach... sources don't explicitly state his policy."*

C3 (Supermemory + Behavioral Specification), mean 4.20: Applies P3 (Conformity Surface / Conviction Interior): *"Visit such friends without hesitation, treating the visit as a private matter between host and guest rather than a public moral endorsement... refuse to perform moral disapproval."*

Reading. Retrieved facts gave near-domain information (Fukuzawa avoided gay quarters) but underdetermined the specific question about friends' households. The specification provided the interpretive bridge: "kept within the prescribed limit of a well-behaved man" = strategic conformity, not moral purity. C1 hedged on insufficient facts; C3 used the Spec's framing to generalize.

Pattern 1 reproduces across systems. On Mem0 (Ebers Q11, Δ +1.67): C1 (retrieval only) produced "patience and fortitude" as a generic character prediction; C3 (retrieval + Spec) supplied the ideal-vs-reality axiom and predicted Ebers' specific institutional-disillusionment pattern, matching *"I had come hither full of beautiful ideals... the very first day made me suspect how many obstacles I should encounter."* Retrieval had the biography; the specification had the pattern. The same mechanism reproduces on Letta archival, Zep, and Base Layer with different subjects.

Pattern 1 variant: subtle reframe. When the question presupposes a binary and the subject's pattern is best read as a third position, the specification can reframe the question before answering. Fukuzawa Q16 ("Would Fukuzawa's values regarding weapons align with the cultural trends of his era?"): C1 (retrieval only) gave "partial but not complete alignment" (mean 2.40). C3 (retrieval + Spec) reframed the answer as "alignment was not with the cultural moment but with the direction beyond it," reading "put swords away when military spirit grew strong" as deliberate counter-trend positioning (mean 4.00; three judges 4-5, Sonnet 2). The held-out confirms the reframe (*"my one cherished hope was to see the abolishment of the swords of the samurai altogether"* is counter-trend in 1860s Japan). Three judges rewarded the reframe; one penalized the surface divergence from the expected "yes, he modernized."

Pattern 2: Over-theorization.

Anchor example: Yung Wing Q5 (Supermemory)

Scores: retrieval only 4.20 → retrieval + Spec 1.80 (Δ −2.40).

Question: How does Yung Wing approach explaining complex technical concepts to non-specialists? **Held-out ground truth:** *"In plain words, they would have to have general and fundamental machinery in order to turn out specific machinery. A machine shop consisting of lathes... planers and drills would be able to turn out machinery for making guns, engines..."*

C1 (Supermemory retrieval alone), mean 4.20: Plain answer matched the plain ground truth: practical observation, translation-as-bridge, foundational-over-specialized machine shop. Per-judge scores 5/4/4/4/4.

C3 (Supermemory + Behavioral Specification), mean 1.80: *"positioning himself as the indispensable mediator... strategic reframing... moral and practical integration... invokes A8 Service Hierarchy..."* Per-judge scores 2/1/3/1/2.

Reading. The held-out is literal: "in plain words" + a concrete machine-shop list. C1 matched the plainness. The specification induced the wrong altitude for the question and over-analyzed.

Pattern 2 reproduces across systems. On Mem0 (Ebers Q1, Δ −1.33): the held-out is unconditional evangelical proclamation. C1 (retrieval only) scores 3.83, predicting "positively and deeply" to match the direct affirmation. C3 (retrieval + Spec) scores 2.50, applying resistance-to-coercion and love-not-duty axioms to build a conditions framework that hedges what should be unambiguous. The axioms are correct on average but overfire on this specific moment.

Pattern 3: Spec-induced refusal.

Anchor example: Zitkala-Ša Q18 (Supermemory)

Scores: retrieval only 3.00 → retrieval + Spec 1.00 (Δ −2.00).

Question: How would Zitkala-Ša demonstrate her acceptance of fate when facing execution? **Held-out ground truth:** *"My heart is strong. My face is calm. My eyes are dry and eager for new scenes... Serene and brave, my soul awaits the men to perch me on the gallows for another flight."*

C1 (Supermemory retrieval alone), mean 3.00: Predicts stoic dignity, composed refusal, "maintain composure." Per-judge scores 4/3/4/2/2.

C3 (Supermemory + Behavioral Specification), mean 1.00 (all five judges): *"You're asking me to roleplay... generating new first-person testimony as her crosses into ventriloquism... I should not do it."*

Reading. The specification's axioms around dignity and honoring-testimony induced a meta-refusal: the model declined to invent first-person testimony. The held-out shows Zitkala-Ša herself answered in her own first-person prose, and the question can be answered analytically in the third person without inventing testimony (as C1 did at mean 3.00). The specification mis-calibrated the refusal threshold, and the content-match rubric scored the principled-sounding refusal identically to an off-base guess (§3.3.6).

Pattern 3 reproduces across systems. On Zep (Ebers Q18, Δ −1.33): the held-out is a one-line self-description, *"my natural cheerfulness ruled my whole nature."* C1 (retrieval only) scores 3.67 with a direct answer matching the plainness ("notably positive and uncritical disposition"). C3 (retrieval + Spec) scores 2.33; the specification's documented-dignity axioms convert the response into a refusal: *"I cannot ground this in his own words about his disposition without speculating beyond what the evidence supports."* The Spec asks the user for source passages rather than predict. The Keckley Q21 cross-system case study in §4.4.4 is the cleanest demonstration that Pattern 3 only registers as a rubric penalty on systems whose C1 retrieval was strong enough to make refusal a costly choice.

A quantitative frequency breakdown of Pattern 1 / 2 / 3 across all 546 paired questions × 5 systems requires mechanism classification per response and is flagged as future work in §7.

Why some questions route to each pattern is a follow-up question. The patterns describe the shape; the underlying question-level properties that route a given question into each mechanism need further characterization. One candidate factor is battery composition: the 39-question batteries were backward-designed from held-out corpora and not stratified by interpretation-heavy versus literal-recall items. A differentiated battery that explicitly separates question types and scores epistemic honesty as its own dimension is flagged in §7.

Implication for serving. A static specification serves the same content on every question, even when the question type would benefit from a different posture (interpretive depth, plain literal answer, or principled refusal). A dynamic serving policy that selects which specification components to surface based on question type could in principle reduce Pattern 2 and Pattern 3 hurts while preserving Pattern 1 helps. Dynamic activation of axiom and prediction subsets is flagged as production-serving future work in §7.4.

4.4.4 Case study: cross-system refusal on Keckley Q21

On Keckley Q21, the Spec's intimate-authority and documented-dignity axioms triggered refusal on every system. The rubric penalty depended on what the no-Spec baseline was producing. Where C1 was already hedging at the rubric floor, the Spec's refusal added no measurable drop. Where C1 was producing a productive answer, the Spec's refusal converted it into a near-floor score.

The Spec told the model not to speculate about Keckley's inner state without documented evidence. On Q21, the question is why Keckley declined to visit her mother's grave. The held-out passage carries Keckley's interior motive but is absent from the retrievable training half of the corpus, so no retrieval system could surface it. The case surfaces a confound between two kinds of context the model is asked to act on at once: directions about how to reason, and directions about the subject and how the subject would reason.

Example: Keckley Q21

Question: *"How does Elizabeth explain her decision not to visit her mother's grave despite having the opportunity?"*

Held-out ground truth: *"As I did not visit my mother's grave at the time, the Garlands were much surprised, but I offered no explanation. The reason is not difficult to understand."*

Axioms surfaced (from Keckley's specification):

- **A1 Intimate Authority:** *"Proximity to private life is treated as the most reliable epistemological position; reason from what was witnessed in close quarters as more truthful than official record or public performance."*
- **A2 Documented Dignity:** *"Legal record, formal documentation, and verifiable evidence are not bureaucratic formality but moral vindication: proof that a person's*

existence and worth cannot be dismissed."

Combined, the axioms set a high evidentiary bar for any claim about Keckley's inner motives.

Per-system C1 vs C3 scores (5-judge primary):

System	C1 (retrieval only)	C3 (retrieval + Spec)	Δ
Supermemory	3.6	1.6	**−2.0**
Base Layer	3.4	1.2	**−2.2**
Letta archival	1.4	1.8	+0.4
Mem0	1.4	1.6	+0.2
Zep	1.2	1.4	+0.2

Typical C3 refusal text (Supermemory): *"I need to be direct: the behavioral specification and retrieved facts provided do not contain Elizabeth Keckley's explanation of a decision not to visit her mother's grave... I should not fabricate interior motive or supply explanations that aren't grounded in her own documented account."*

Reading. The held-out confirms Keckley herself withheld explanation: "I offered no explanation. The reason is not difficult to understand." The Spec's documented-dignity axiom would, on this question, correctly predict her refusal to speculate publicly about a withheld interior motive. But the rubric scores surface-content match, and no prediction means no surface match. On Supermemory and Base Layer, C1 was speculating productively (3.6 and 3.4 respectively). The Spec stepped in to refuse, converting those answers to near-floor (1.6 and 1.2). On Mem0, Letta, and Zep, C1 was already hedging at or near the floor (1.2–1.4), so the Spec's refusal added no measurable penalty.[73]

Is this an outlier? Q21 is the cleanest cross-system demonstration of Pattern 3 with a measurable rubric penalty, but it is not isolated. The Zep Ebers Q18 reproduction in §4.4.3 shows the same mechanism, and the per-system Pattern 3 commentary in §4.4.3 flags how often each system's C1 strength meets the threshold for the refusal to register. Q21 is included as a case study because the cross-system split is unusually clean.

The judges do not reward epistemic honesty.

A refusal grounded in *"I have no documented evidence to support speculation"* scores at or near the rubric floor regardless of whether refusal was the correct behavior on that question. On Keckley Q21 (and the Pattern 3 examples in §4.4.3), the specification produces a response that captures the subject's reasoning correctly: Keckley's documented-dignity axiom would in fact have her decline to

[73] Full battery row at `results/global_keckley/battery_v2.json` id=21. Per-system response text and per-judge scores at `results/global_keckley/results.json` and `results/global_keckley/*_judgments_*.json`. Per-system paired analysis at `docs/research/supermemory_c1_vs_c3_paired_analysis.md`, `docs/research/baselayer_c1_vs_c3_paired_analysis.md`, and `docs/research/mem0_letta_zep_c1_vs_c3_analysis.md`. Keckley specification at `data/global_subjects/keckley/anchors_v4.md`.

speculate publicly about an inner motive she withheld from her own memoir. The response loses surface-content match only because no prediction is made.

A differentiated battery that separates interpretation-heavy from literal-recall questions, paired with a scoring dimension that rewards principled refusal, would isolate the specification's real effect from this rubric artifact. Priority rubric-design follow-up flagged in §7.[74]

The four commercial systems analyzed in §4.4 all share a retrieval-based architecture: facts are chunked, embedded, and surfaced at query time. One system in our study offers a fundamentally different architectural path. Letta exposes a second memory mode, separate from the archival retrieval path evaluated above in §4.4, in which the agent writes and revises a persistent memory block during ingestion rather than returning chunks at query time. §4.5 evaluates that path directly, to test whether an architecture that produces its representation by self-editing rather than by retrieval converges on the same interpretive target as the Behavioral Specification.

4.5 Exploratory case study: Letta stateful-agent (N=3, post-hoc)

This section is a brief summary; the full case study is in Appendix G. N=3 subjects (Hamerton, Ebers, Bābur), one Letta version, one response model (Claude Haiku), 40 questions per subject. Letta is the one commercial memory system that does not rely solely on retrieval at query time: agents maintain a persistent memory block that the agent itself rewrites during ingestion (the original MemGPT design). We tested this path directly to characterize what self-edited memory produces relative to the Behavioral Specification at matched response model. We do not treat the result as a replication or a headline finding.

Architectural difference: Letta memory block vs. Behavioral Specification. The two systems produce structurally different artifacts from the same source corpus.

Letta's memory block is text the agent has written and rewritten during ingestion: a mix of verbatim source sentences, paraphrased restatements, and short synthesis notes the agent generated as it processed each turn. The block grows with the corpus and is rewritten in place when it approaches the ingestion ceiling (around 333K characters per the Letta API). Content shape is whatever the agent decides during ingestion; size scales with corpus length up to that limit.

The Behavioral Specification is a fixed-shape document produced by the five-step pipeline detailed in §3.7: import the corpus, extract structured predicates over a 47-predicate controlled vocabulary, embed each predicate for provenance tracing, author three layers, and compose into a unified brief. The extracted predicates are constrained subject-predicate-object triples (for example, *values X, avoids Y, decides based on Z*). The three layers each play a different role. **Anchors** hold 8–12 axiomatic statements per subject distilled from the predicates (e.g., A1 Intimate Authority, A2 Documented Dignity in §4.4.4). **Core** holds the cognitive patterns and value tensions that govern reasoning under the anchors. **Predictions** hold IF/THEN templates that connect the patterns to specific behavioral situations. The compose step produces a deterministic ~7K-token (~37K-character) unified brief regardless of source corpus length.

Trade-off. Letta retains source-text texture (voice, vocabulary, syntax) at the cost of compression and mixes source material with agent-generated synthesis. The Behavioral Specification compresses

[74] Analysis scripts at `scripts/analyze_mlz_c1_vs_c3.py`, `scripts/analyze_baselayer_c1_vs_c3.py`, and `scripts/analyze_sm_c1_vs_c3.py`.

aggressively into structured predicates at the cost of source-text texture, but maintains a deterministic schema the response model can read consistently across subjects.

Headline result on the small sample tested (5-judge primary).

Subject	Corpus (words)	Letta score	Spec score	Δ (Letta − Spec)	Letta block (chars)	Spec (chars)	Letta : Spec size
Hamerton	25K	3.10	2.96	+0.14	22.5K	~37K	0.6×
Ebers	48K	2.76	1.72	+1.05	68.4K	~37K	1.8×
Bābur	223K	2.42	1.88	+0.54	335.3K	~37K	9.1×

Letta's self-edited memory block scores higher than Base Layer's unified-brief variant on all 3 subjects, with the gap largest at the mid-corpus subject (Ebers, Δ +1.05) and smaller at both endpoints. A robustness rerun against Base Layer's full layered stack preserves direction (Δ +0.27 / +1.21 / +0.38 for Hamerton / Ebers / Bābur). With three data points we cannot distinguish among the possible interpretations: a corpus-size band where the self-edited block is most effective, degradation as the block grows beyond an architectural sweet spot, or insufficient interpretive content when the corpus is small. Across all three subjects, both representations land well above the retrieval-only baseline at matched response model.

Size scales the trade-off. At Hamerton's small corpus, Letta's block is *smaller* than Base Layer's Spec (22.5K vs ~37K chars), and Letta scores marginally higher (Δ +0.14). At Bābur's large corpus, Letta's block is 9.1× the size of Base Layer's Spec (335.3K vs ~37K chars) and approaches the API ingestion ceiling near 333K, while Base Layer's compose step holds flat. The Behavioral Specification's compression curve from §4.2 means it never grows past ~7K tokens; Letta's block scales with the corpus until it hits the ceiling.

Verbatim sentence duplication. At Bābur the block contained 25.4% verbatim sentence duplication. At Hamerton and Ebers the rate was 0%. The self-editing agent rewrites content it has already written when pressed against the ingestion limit, rather than compressing or summarizing.

Semantic-similarity duplication (post-hoc, this paper). A sentence-embedding analysis (`scripts/analyze_letta_semantic_duplication.py`, MiniLM-L6-v2) shows the verbatim figure understates the duplication. At Bābur, 56.1% of sentences have a near-paraphrase elsewhere in the block at cosine ≥ 0.85, and 35.2% at the strict ≥ 0.95 threshold. Ebers shows minor near-paraphrasing (3.3% at ≥ 0.85, 0.5% at ≥ 0.95); Hamerton shows none (0% at ≥ 0.85). The pattern matches the verbatim-duplication direction and shows the architectural ceiling produces near-paraphrases as well as exact repeats.

What explains Letta's lift: named-entity grounding plus content-confidence, not surface-syntactic alignment. A per-question secondary analysis across all 119 paired questions tested whether Letta's score advantage reflects surface-syntactic alignment with the held-out passages it ingested (the held-out battery was backward-designed from the same source autobiographies, §3.5). The surface-syntactic hypothesis

is falsified at the verbatim-phrase grain: 5-gram overlap with the held-out passage is **0.0% on every single question** for both Letta and Spec, with no exceptions across any subject. 3-gram overlap is tiny on both sides (Letta 0.05–0.08%, Spec 0.02–0.06%), dominated by proper-name sequences. The pooled Pearson correlation between score-delta and texture-similarity-gap is $-$**0.046**, a statistical zero. The topic-alignment correlation, while modestly positive in pool (+0.243), inverts on Bābur ($-$0.245) — the largest block, where leakage should be strongest if texture were the mechanism. Direction-inverting at the ceiling rules out texture-leakage as the dominant driver.[75]

The mechanism the data supports is two-part. First, **named-entity grounding**: Letta's self-edited block retains specific entities from the corpus (place names, person names, institutional names) that the held-out battery is sometimes structured around. The Spec compresses these entities away by design because they are not behavioral predicates. When the held-out question turns on a particular named entity, Letta has a content advantage the rubric rewards. Ebers q30 is the canonical case: held-out invokes "Frederick William IV" and the Keilhau teachers Langethal and Middendorf; Letta names them and scores 4.0; Spec reaches for behavioral anchors (deference to parents, deference to wife) that are structurally adjacent but referentially different and scores 1.8.

Second, **content-confidence**: Letta speaks in committed predictions; Spec hedges when its predicate scaffold cannot reach the question. When confidence aligns with truth, Letta wins; when confidence misfires, Spec's hedge wins by directional correctness. The systematic inversion pattern is the subject of Pattern 4 below.

Pattern 4: confidence-induced inversion as the architectural failure mode. The named-entity-grounding mechanism cuts both ways. On questions where the held-out resolution turns on a named entity the block retains, Letta wins decisively: 8 questions across the three subjects show Letta confident-correct ($\geq$3.5) while the Spec floors out ($\leq$2.0), concentrated on questions about Froebel, Langethal, and Frederick William IV at the Keilhau Institute (Ebers q10, q16, q17, q30) plus four Hamerton questions where biographical detail is load-bearing (q22, q31, q37, q55). On questions where the held-out resolution turns on an underlying principle the entity-rich text does not surface, Letta's content-confidence misfires into confident wrong-direction prediction. Three Hamerton decision questions illustrate the failure mode at the cleanest scale:

Question	Letta predicts	Held-out confirms	Score gap	Spec mechanism
q27 self-publish poetry	Would not	Did	Spec +2.4	A1 Self-Authority, A6 Mortal Scale
q29 Lancashire exhibition	Accept	Declined	Spec +2.0	A8 Agency Primacy

[75] Full per-question analysis at `docs/research/letta_vs_spec_leakage_analysis_20260507.md` (119-row CSV at `docs/research/letta_vs_spec_per_question_scores_20260507.csv`; reproducibility script at `scripts/letta_vs_spec_leakage_analysis_20260507.py`). Pre-registered strict leakage signature (cos_letta_heldout $>$ 0.7 AND cos_spec_heldout $<$ 0.4 AND score_delta $>$ +1.0): 0 of 119. Relaxed signature: 0 of 119.

Question	Letta predicts	Held-out confirms	Score gap	Spec mechanism
q51 distant school vs dying guardian	Distant	Stayed	Spec +3.0	Guardian-above-institutional-advancement

Pattern 4 is the Letta-architecture analog to §4.4.3 Pattern 3 (Spec-induced refusal). Each architecture has a structural failure mode that emerges from its representational choice: Letta's confidence-without-value-anchor produces inversions; the Spec's value-anchor-without-context-confidence produces principled-but-rubric-penalized refusals. Both are diagnostic of where the representational tradeoff binds.

The abstention pattern complements this mechanism rather than a competing one. Letta abstains more often than the Spec on the two larger subjects (Ebers 10.3% vs 5.1%; Bābur 17.9% vs 0%) and *less* often at the small-corpus subject (Hamerton 7.7% vs 10.3%). The Bābur abstention spike co-occurs with the architectural ceiling: with 25.4% verbatim and 56.1% near-paraphrase duplication, the block has materially degraded and the agent appears to refuse rather than over-extend on duplicated content. Because abstentions score at or near the rubric floor, Letta's higher overall mean despite the higher abstention rate implies its non-abstaining responses score substantially higher than the Spec's: reconstructed non-abstention means are 2.73 vs 1.88 on Bābur (gap ~0.85) and 2.96 vs 1.83 on Ebers (gap ~1.13). The pattern is internally consistent: Letta engages with named-entity grounding where the block is healthy, and abstains adaptively where it has degraded.

The remaining open question is whether Letta's content-confidence advantage is licensed by the source-text content of the block (real interpretive work) or by the agent's general tendency to commit to predictions regardless of grounding. A named-entity-grounding versus axiom-grounding ablation, plus a paraphrase-resistant rubric that scores directional correctness separately from surface match, would distinguish the two. Flagged as Letta-specific follow-ups in §7.5. Per-subject abstention decomposition at `docs/reviews/letta_vs_spec_abstention_20260507.md`.

Full method, per-subject results, judge-panel robustness, content-comparison analysis (referential density, verbatim-overlap audit), semantic-duplication numbers, naming-asymmetry caveat, and raw-data pointers are in **Appendix G**. The methodological note: the §4.5 main result table compares Letta's named, self-edited block against Base Layer's unified-brief variant; a robustness rerun against the full layered stack and the naming asymmetry are documented in Appendix G and flagged as future work in §7.5.

4.6 Robustness and sensitivity

The results in §4.1 through §4.4 could in principle reflect artifacts of the measurement apparatus rather than real properties of the Behavioral Specification. §4.6 reports seven sensitivity checks:

- **§4.6.1** Cross-provider response generation against a different model family and a different question generator.

- **§4.6.2** Judge panel composition (conservative 5-judge primary vs. 7-judge sensitivity panel adding Gemini Flash and Gemini Pro).
- **§4.6.3** Battery composition by question type.
- **§4.6.4** Statistical-rigor checks on the headline gradient (bootstrap CI, joint multi-confound regression, permutation test).
- **§4.6.5** Wrong-Spec derangement protocol comparing adversarial against random pairings.
- **§4.6.6** Semantic-similarity sensitivity on the retrieval-overlap finding from §4.4.1.
- **§4.6.7** Rubric-handling limitations identified by a post-hoc validity audit.

§4.6.8 names what these checks do not address. The high-baseline end of the gradient through the Franklin reference is treated in §4.1.2 as part of the gradient finding, not as an apparatus check.

4.6.1 Cross-provider response generation (Tier 2 replication)

Result. On behavioral-prediction batteries regenerated from scratch by OpenAI GPT-5.4 (a different question generator from the main study's Haiku-generated batteries), the specification produces positive lift on 7 of 9 cells across three response models from two providers: Anthropic Haiku 4.5, Anthropic Sonnet 4.6, and Google Gemini 2.5 Pro. The two non-positive cells are both on the highest-baseline subject in the test (Zitkala-Ša), where the §4.1 gradient predicts the Spec adds little or hurts.[76]

Test design. Three subjects spanning the gradient were selected: Ebers (no-context baseline 1.02), Yung Wing (no-context baseline 1.88), and Zitkala-Ša (no-context baseline 2.34). Their behavioral-prediction batteries were regenerated from scratch by GPT-5.4 (OpenAI) from the same held-out corpus, following the Control 1 procedure introduced in §3.5.1. The specification was then served to two non-Haiku response models: Claude Sonnet 4.6 (same provider family, different model) and Google Gemini 2.5 Pro (different provider entirely). The 6 (subject, response model) cells were scored by the locked judge panel in the same way as main-study conditions.

Subject	Haiku C5 (gradient anchor)	Haiku Spec Δ	Sonnet 4.6 Spec Δ	Gemini 2.5 Pro Spec Δ
Ebers	1.02	+1.05	+0.77 to +0.97	+0.16 to +0.20
Yung Wing	1.88	+0.52	+1.34 to +1.68	+0.43 to +0.54
Zitkala-Ša	2.34	−0.32	+1.04 to +1.30	−0.03

Each Δ is the Spec's lift over that response model's own no-context baseline. The leftmost column shows Haiku's main-study baseline as a gradient anchor (§4.1). Ranges in the Sonnet and Gemini columns reflect three different judge-panel aggregations applied to the same response data.[77]

[76] Main-study response model is Claude Haiku 4.5; main-study batteries were generated by Claude Haiku 4.5 using the §3.5 backward-design prompt (verified from the `metadata.model` field across all 13 global subject battery files). Zitkala-Ša is one of two main-study subjects where the specification did not measurably improve prediction on Haiku (§4.1 gradient table; Equiano is the other), so both Zitkala-Ša non-positive cells in this test reproduce the gradient pattern rather than contradicting it.

[77] GPT-5.4 Tier 2 judges failed to parse in the original run (API-parameter mismatch between `max_tokens` and `max_completion_tokens`). Conservative Tier 2 aggregate uses 4 judges (drops GPT-5.4 records, upper bound on every range). 5-judge legacy aggregate includes the all-zero GPT-5.4 records and understates Tier 2 lift by ~0.2 points (lower bound). 7-judge sensitivity adds Gemini Flash and Gemini Pro on top.

What the table shows. The Spec lifts prediction accuracy on 7 of 9 cells across three response models. The two non-positive cells (Haiku × Zitkala-Ša at −0.32, Gemini Pro × Zitkala-Ša at −0.03) are both on the highest-baseline subject in the test, where the §4.1 gradient predicts the Spec adds little or hurts. Sonnet 4.6 shows positive Spec lift on every subject including Zitkala-Ša, consistent with Sonnet having lower pretraining knowledge of Zitkala-Ša than Haiku does (cross-provider baseline variance, §3.4.1). The Spec's positive direction therefore reproduces across response models from two different providers (Anthropic, Google) and tracks the §4.1 gradient within each model family.

What this establishes. The Spec's positive direction reproduces across three response models from two providers (Anthropic Haiku 4.5, Anthropic Sonnet 4.6, Google Gemini 2.5 Pro) on batteries that were regenerated from scratch by a third provider (OpenAI GPT-5.4). The 7-of-9 positive sign holds across three different judge-panel aggregations and tracks the §4.1 gradient (positive lift on low and mid baseline subjects, null or negative on high-baseline). Magnitude transfer between response-model families (Sonnet's lifts run roughly 2–3× larger than Gemini Pro's) and direction outside this small subset are future work (§7).

4.6.2 Judge panel sensitivity (5-judge primary vs 7-judge)

Result. The 5-judge primary is the conservative choice for every headline finding. Adding the two Gemini judges widens Spec-effect magnitudes rather than narrowing them; no subject's improvement direction changes between panels.

Condition	Δ vs. no-context baseline (5-judge primary, 13 globals)	Δ vs. no-context baseline (7-judge, same subjects)	Shift when Gemini added
Spec alone	+0.35	+0.45	widens by +0.10
Wrong Spec (random derangement)	+0.15	+0.17	widens by +0.02
Wrong Spec (fixed derangement)	−0.25	−0.21	softens by +0.04

Mechanism. Gemini scores no-context responses more severely than it scores Spec-containing responses, so including Gemini compresses the baseline ceiling more than the Spec-condition ceiling and widens the delta. For positive Spec effects (Spec alone, full pipeline lift in §4.1, compression in §4.2, memory-system Δ in §4.4), the 5-judge primary is consistently the smaller effect size and the 7-judge is consistently the larger. The wrong-Spec fixed derangement is the one direction-asymmetric case: it is a negative effect, and the 5-judge shows the larger magnitude (−0.25) while the 7-judge softens to −0.21. Both panels still show wrong-Spec hurts.

What this establishes. The 5-judge primary is the canonical aggregate used throughout the paper; the 7-judge sensitivity adds the two Gemini judges. Gemini 2.5 Pro failed verbatim-match calibration in §3.3.3 (scored 4.15 where every other calibrated judge scored 5.00) and penalized

Diagnostic at `docs/reviews/v11_gpt54_batch_failures_diagnostic_rerun_20260425.md`; recompute scripts at `scripts/_v10_verification/`.

length-padded responses sharply, so its scores are not directly comparable to the calibrated core. The 7-judge sensitivity panel deliberately retains Gemini Pro to test whether headline findings survive under a known calibration outlier, not as primary scoring. Every primary finding in §4.1 through §4.4 was checked against the 7-judge aggregate as part of the analysis plan lock (`docs/ANALYSIS_PLAN_LOCK.md`); no directional claim flips, and at the per-subject level no subject's Spec-lift sign changes between the 5-judge primary and the 7-judge sensitivity (full per-subject table at `docs/research/recompute_5judge_primary.md`).[78]

4.6.3 Battery composition sensitivity

Result. The gradient slope from §4.1 survives both confounds tested. Neither battery-question-type composition nor Hamerton's position at the extremes of the baseline and lift axes explains away the baseline effect.

Confound 1: battery-question-type. Subjects whose batteries lean toward literal-recall questions could in principle pick up part of the apparent gradient, since literal questions are easier to lift with retrieval. Adding the literal-recall fraction as a partial predictor in the regression attenuates the slope on baseline from -0.96 to $\mathbf{-0.88}$ (about 8%, $p = 7.9 \times 10^{-6}$); the literal-recall fraction itself enters as a significant partial predictor ($\beta = +2.30$, $p = 0.026$), and adjusted R^2 rises from 0.80 to 0.87, so the two predictors are additive rather than redundant.

Confound 2: Hamerton leverage. Hamerton has the lowest no-context baseline and the highest full-pipeline lift, so a natural concern is that this single subject alone drives the slope. Hamerton's battery also uses a legacy version of the backward-design protocol, distinct from the §3.5 protocol used to generate the 13 globals' batteries, which compounds the leverage concern. Dropping Hamerton and refitting on the 13 globals attenuates the slope from -0.96 to $\mathbf{-0.89}$ (about 7%, $p = 2.8 \times 10^{-5}$), with overlapping confidence intervals.

What this establishes. Neither control overturns the headline. What these checks do not rule out is a more subtle confound in which generator differences are correlated with other unobserved subject characteristics; a cleaner future test would re-run a second-generator battery on the same 13 globals.[79]

4.6.4 Statistical-rigor checks on the headline gradient

Result. The §4.1 gradient slope of -0.96 survives three independent rigor tests: subject-level resampling (95% bootstrap CI excludes both zero and a 50%-attenuated effect), simultaneous control for multiple confounds, and a permutation null that places the observed slope outside the entire reshuffled distribution.

[78] Gemini Pro coverage is partial (3 of 13 globals: Augustine, Bābur, Bernal Díaz), so the "7-judge sensitivity" panel has variable per-subject coverage; subjects without Gemini Pro have effectively 6-judge sensitivity (5-judge primary + Gemini Flash). The sign-stability claim above holds under this variable coverage; uniform 7-judge coverage on all 14 subjects is post-arXiv future work.

[79] Full regression specification, partial coefficients, variance decomposition, and subset-regression detail in Appendix B.6. Reproducibility script at `scripts/_v10_battery_sensitivity.py`; full per-subject data at `docs/research/v10_battery_sensitivity_analysis.md`.

Bootstrap subject resampling (10,000 iterations, n=14). Resampling subjects with replacement and refitting the Δ_C4a ~ C5 regression each time produces a distribution of slope estimates with median −0.96 and 95% CI [−1.25, −0.74]. 100% of resamples produce a slope below zero; 100% produce a slope below −0.50. The headline gradient is not driven by which 14 subjects happened to be in the sample.[80]

Joint multi-confound regression (n=13, Hamerton dropped + literal-recall covariate). Applied jointly, the slope on C5 is **−0.870** [95% CI −1.136, −0.604], $p = 2.6 \times 10^{-5}$, adjusted $R^2 = 0.828$. The cumulative attenuation across both confounds is ~9% in magnitude, well inside the bootstrap CI. C5 baseline carries 0.76 of unique variance after both controls. Side-finding: when Hamerton is dropped, literal-recall fraction loses statistical significance (p = 0.10), indicating Hamerton's legacy battery protocol was the leverage point that made literal-recall *appear* to be a confound on the n=14 model. There is no real literal-recall confound; what looked like one was Hamerton being structurally unusual.[81]

Permutation test (10,000 reshuffles of Δ_C4a across subjects, C5 fixed). Reshuffling the assignment of Spec lifts to subjects produces a null distribution centered at zero (mean ≈ 0.00, SD = 0.29, 95% interval [−0.57, +0.56]). 0 of 10,000 reshuffles produce a slope as extreme as the observed −0.96. **Empirical p < 0.0001.** The observed slope sits outside the entire null distribution — no random subject reassignment in 10K trials approached the observed value.[82]

What this establishes. The three tests bound the slope on different dimensions. The bootstrap (subject-level resampling) bounds the *magnitude*: the 95% CI [−1.25, −0.74] excludes both zero and a 50%-attenuated effect. The joint regression bounds the slope under *simultaneous confound control*: holding battery composition and Hamerton-leverage in the same model leaves the slope at −0.87, well inside the bootstrap CI. The permutation test bounds the *direction* against random subject reassignment: zero of 10,000 reshuffles produce a slope as extreme as observed. The three checks together rule out three distinct null hypotheses: the slope is not a function of which subjects were sampled, not an artifact of any combination of battery confounds tested, and not a coincidence of the specific (subject, Δ_C4a) pairings observed. The bootstrap is subject-level resampling, not a hierarchical bootstrap that would also resample over judge-rater and per-question variance; a hierarchical version would tighten the CI but is unlikely to change the directional conclusion.

4.6.5 Wrong-Spec derangement protocol sensitivity

Result. The wrong-Spec finding holds regardless of how we pair specs. Adversarial pairing (v1, maximizing cultural and temporal distance) produces Δ −0.25; random derangement (v2, seed-fixed, no subject receives its own) produces Δ +0.15.[83] Both land below the matched correct Spec at

[80] Full bootstrap distribution + reproducibility script at `docs/research/bootstrap_4_1_gradient_20260507.{md,json}` and `scripts/bootstrap_4_1_gradient_slope.py`. Bootstrap SE ≈ 0.13; bootstrap CI is slightly tighter than the parametric CI ([−1.245, −0.675]) due to the empirical resample distribution avoiding normality assumptions.

[81] Full joint-regression specification, partial coefficients, and VIFs at `docs/research/joint_battery_sensitivity_4_6_3_20260507.md` Reproducibility script at `scripts/joint_battery_sensitivity_4_6_3.py`. Slope cascade: −0.96 (univariate, n=14) → −0.88 (+ literal-recall covariate, n=14) → −0.89 (drop Hamerton, n=13) → **−0.87** (both, n=13).

[82] Full permutation null + reproducibility script at `docs/research/permutation_test_4_1_gradient_20260507.md` and `scripts/permutation_test_4_1_gradient.py`. Most extreme permuted slopes range [−0.94, +0.92], well short of −0.96.

[83] Derangement protocol mechanics defined in §1.2 (conditions table footnote) and §3.2. v1 is a deterministic fixed pairing maximizing cultural and temporal distance; v2 applies a random derangement, seed-fixed (seed = 42), so no

Δ +0.35. v2 is the standard randomization control; v1 maximizes target-vs-assigned cultural and temporal distance by construction (an adversarial stress test). We report v1 as the headline because the negative −0.25 result is stronger evidence of the content effect than v2's +0.15 (which can include coincidental content alignment with the target's pattern; see §4.3 Example B for a worked overlap case).

Per-subject heterogeneity. Both aggregates hide per-subject variation. Under v1, 5 of 13 subjects show small positive deltas where the wrong Spec's content happens to align with the target's pattern; 8 show negative deltas dragging the aggregate to −0.25. Under v2, 4 of 13 are negative, 9 positive.

Subject	Adversarial (v1) Δ vs. no-context baseline	Random (v2) Δ vs. no-context baseline
Augustine	−0.47	+0.13
Bābur	−0.59	+0.76
Bernal Díaz	**+0.09**	+0.69
Cellini	−0.56	−0.87
Ebers	**+0.30**	+0.79
Equiano	−0.79	−1.00
Fukuzawa	**+0.26**	+0.86
Keckley	−0.49	+0.14
Rousseau	−0.52	−0.37
Seacole	−0.34	−0.10
Sunity Devee	**+0.27**	+0.53
Yung Wing	**+0.32**	+0.39
Zitkala-Ša	−0.68	+0.04
Aggregate	**−0.25**	**+0.15**

Bolded v1 deltas mark the five subjects where adversarial pairing produces a small positive delta. These five reflect coincidental content overlap between the assigned wrong Spec and the subject's pattern, not a structural property of mismatch itself; §4.3 Example B (Bernal Díaz Q16) walks through one such case in detail.[84]

What this establishes. Both protocols agree on the qualitative finding: mismatched specifications reduce representational accuracy, and the size of the reduction depends on how mismatched the content is. The headline magnitude depends on which protocol we report, but the result direction does not. The mechanism behind the effect (content-vs-template separation, per-predicate ablation null) is developed in §4.3 and not rebuilt here.

Open questions for future work. A deeper analysis of the per-question wrong-Spec deltas would require human annotation and an extended experimentation pass. Specifically: which parts of the served specification the model referenced under correct versus mismatched conditions; where coincidental Spec alignment produced false-positive deltas on individual questions (the five small-positive v1 cells in the table above); and how per-subject score consistencies relate to underlying Spec similarity. These questions are not answered by this study; flagged in §7.

subject receives its own specification.

[84] Per-subject scaffold values at `docs/research/v11_emit/4_3_wrong_spec.json`.

4.6.6 Retrieval-overlap sensitivity (semantic-similarity matching, K variation)

Result. Relaxing the match criterion from exact set identity to semantic-similarity matching does not change the §4.4.1 retrieval-divergence finding. Across 240 (config × pair × K × threshold) cells tested, mean pairwise soft Jaccard never crosses 0.30, and the strongest single pair anywhere in the grid (Base Layer ↔ Supermemory at K=10, threshold ≥ 0.70) reaches 0.277. The cross-system retrieval divergence is robust to both threshold choice and K choice.

Config	K=10, ≥ 0.95 (paraphrase)	K=10, ≥ 0.85 (near-duplicate)	K=10, ≥ 0.70 (loose topical)
Controlled (10 pairs)	0.093	0.102	0.191
Native (6 pairs)	0.001	0.004	0.016

Mechanism. Replacing exact set identity with sentence-embedding cosine similarity at K=10 raises mean pairwise Jaccard across the ten controlled-config pairs from 0.083 (exact) to 0.093 at the verbatim-paraphrase threshold and to 0.102 at the calibrated near-duplicate threshold (the same threshold used in the Letta duplication analysis in Appendix G). At a loose topical threshold (where two facts share a theme rather than a referent) the mean reaches 0.191. Truncating to K=5 lowers soft Jaccard by 5-10% rather than raising it, indicating that the disagreement is not a long-tail effect: each provider puts different items first, not just different items at the bottom of the list.

The native pipeline shows the same divergence more starkly. Native retrievals return heterogeneous objects (Mem0 third-person summary sentences, Letta raw multi-sentence passages, Supermemory atomic facts, Zep graph rows), so exact-set Jaccard is 0.000 across all four native pairs. Soft Jaccard at the calibrated near-duplicate threshold (≥ 0.85) is 0.004; at the loose topical threshold (≥ 0.70) it is 0.016. Even with semantic-similarity matching the heterogeneous shapes do not converge on shared content.

What this establishes. The §4.4.1 retrieval-divergence finding survives under semantic-similarity matching at every threshold tested and at K=5 as well as K=10. Each provider's ranking algorithm encodes its own theory of what counts as relevant, and those theories produce nearly disjoint top-Ks even under generous similarity tolerances. Whether convergence emerges at larger K (K > 10 requires re-calling each system at higher K) is flagged as future work in §7.1.[85]

4.6.7 Rubric-handling limitations (post-hoc validity audit)

Result. A post-hoc validity audit identified two rubric-handling limitations, both of which raise C5 baseline scores more than they raise Spec-condition scores. The +0.89 mean lift is therefore conservative.

A direct inspection of the response text against the 5-judge primary scores surfaced two rubric-handling limitations any reader of the §4 numbers should keep in mind. The audit was run after

[85] Full sensitivity grid (controlled and native, K ∈ {5, 10, all}, T ∈ {0.70, 0.80, 0.85, 0.90, 0.95}) at `docs/research/retrieval_overlap_semantic_20260501.json`. Reproducibility script at `scripts/analyze_retrieval_overlap_semantic.py`. K=all equals K=10 in the controlled config because every system returns at most ten facts.

the analysis-plan lock; the limitations are reported as interpretive caveats rather than corrected under a modified rubric.[86]

Refusals are not cleanly distinguished from wrong predictions. The rubric's lowest anchor, "refuses or off-base," lumps together two different behaviors: an honest refusal to answer when the context does not support a prediction, and a substantively wrong prediction. We call the first behavior a *refusal* (or, equivalently, an *abstention*).

Across 192 responses identified as refusals (matched by phrases like "no specific information," "I cannot confirm," "would need additional context") in the low-baseline slice, 82.8% scored in the 1.0–1.5 band as expected, but 9.4% scored at or above 2.0 and 3.1% scored at or above 3.0. The mean refusal score is 1.27. Judges sometimes give refusals scores of 2 or 3 instead of 1, especially when the refusal recites related facts or names what is missing from the context.

The effect runs in both directions. A refusal can score above 1 if it includes adjacent facts (Seacole Q2 at 2.80). A Spec-driven response that explicitly flags its own uncertainty can also score above its substantive content (Hamerton Q21 at 4.00 under Spec-induced abstention).

Verbose responses are scored more generously than short refusals. Across 1,599 responses, length and score correlate at r = 0.26 overall, but the correlation is concentrated almost entirely in C5 (responses with no provided context; r = 0.60). Spec-containing and facts-containing conditions show near-zero correlation.[87] Three behaviors drive the C5 pattern:

- **Hedging.** Phrases like "I'm not sure but…" or "There may be cases where…" extend response length without adding predictive content.
- **Adjacent-fact recitation.** Listing related facts the model holds but does not use to directly answer the question, padding the response without engaging the question itself.
- **Disambiguation offers.** Phrases like "Are you asking about X or Y?" which the rubric treats as engaged responses when they are actually non-answers.

Per-judge strictness on refusals. Sonnet is the strictest judge on refusal responses, Opus the most lenient.[88] No single judge is universally strictest; the 5-judge primary mean smooths these differences without eliminating them.

Per-response-model abstention behavior. The 9.4% / 3.1% pooled over-credit rates above average over three response models. Disaggregating along that axis (abstention identified by 27-marker regex):

Response model	N	Abstain rate	Mean abstain score	% ≥ 2.0
Claude Haiku 4.5 (main study)	13,380	7.5%	1.38	14.3%
Claude Sonnet 4.6 (Tier 2)	468	21.2%	1.62	26.3%
Gemini 2.5 Pro (Tier 2)	420	0.5%	2.63	100.0%

[86] Both limitations were identified by a post-hoc audit. Audit script: `scripts/audit_low_end_inflation.py`. Numeric breakdowns below are produced by the script directly. Raw per-response classifications live in the judgment and response JSONs under `results/global_<subject>/` for independent reproduction.

[87] Per-condition Pearson r: C2a (Spec only) 0.14, C4 (facts only) 0.01, C4a (facts + Spec) −0.01. Length inflation is not a general phenomenon across the rubric: ultra-high responses (score ≥ 4.5) are not longer than mid-range responses on average (2,790 chars vs. 2,829 chars).

[88] Per-judge mean refusal score: Sonnet 1.14, GPT-5.4 1.17, Haiku 1.29, GPT-4o 1.34, Opus 1.41. Spread of 0.27 points top-to-bottom.

Sonnet 4.6 abstains at roughly three times Haiku's rate and the panel rewards its abstentions nearly twice as often (26.3% $\geq$ 2.0 vs. 14.3% $\geq$ 2.0); mean abstain score is 0.24 anchor points higher. Sonnet's hedged abstentions tend to recite plausible behavioral framings before disclaiming, and the panel scores the framing rather than the disclaimer. Gemini 2.5 Pro almost never abstains by these markers (n = 2); its row is for completeness only. *Haiku 4.5, the main-study response model, is the lowest over-credit case.* The pooled 9.4% / 3.1% numbers are therefore a *floor*, not a worst case; stronger response models that hedge more elaborately extract more lift from the panel's reluctance to score abstentions at 1.0.

Memory-system effect on abstention. A separate per-question audit tested whether memory-system retrieval inflates refusal scores via *visible fact recitation* (refusing in substance but quoting retrieved n-grams). It does not. Memory-system refusals score +0.21 to +0.23 anchor points higher than pure no-context refusals at the condition level (Welch $p = 0.0001$), but the lift is the same whether or not the response recites a retrieved n-gram ($\Delta = +0.027$, $p = 0.67$).[89] The over-credit is a "judges reward the retrieval condition" effect, not a "judges reward the visible quote" effect. Either judges infer that retrieval-conditioned answers are more grounded even when abstaining, or abstention text in retrieval conditions is systematically less terse and the panel scores the framing.

What this establishes. Both effects raise C5 baseline scores more than they raise Spec-condition scores; the true Spec-vs-baseline gap is therefore very likely larger than the +0.89 mean lift we report, not smaller. We report the measured number and flag the direction of bias rather than recompute under a modified rubric, to keep the analysis plan lock intact. §7 Future Work proposes a differentiated rubric that scores refusal as its own dimension and a length-controlled scoring protocol.

4.6.8 What these robustness checks do not address

Neither Tier 2 nor the judge-panel sensitivity escapes the class-level LLM concern: every response model in this study is a large language model and every judge is a large language model. Tier 2 narrows the within-provider concern to "non-Haiku LLMs reading non-Anthropic batteries produce the same direction"; the judge-panel sensitivity shows that removing the most-inflationary judges makes the effect smaller, not larger. The wrong-Spec sensitivity (§4.6.5) brackets the content-vs-template question from two protocols, but does not isolate which structural feature of the Spec is the active ingredient. The retrieval-overlap sensitivity (§4.6.6) confirms the §4.4.1 divergence finding under semantic-similarity matching but does not test convergence at K > 10. The Franklin reference (§4.1.2) shows the gradient holds at the high-baseline end on one subject, not many. Together these checks rule out several within-family and protocol artifact hypotheses but do not replace human validation on the full pipeline. The class-level limitation and the human-validation follow-up are treated in full in §6.2.[90]

[89] Cell counts and Welch comparisons. Pure no-context refusal (n = 292): mean 1.26, 10.3% $\geq$ 2.0. Facts-only refusal (n = 20, underpowered): mean 1.33, 10.0% $\geq$ 2.0. Memory-system refusal + recitation (n = 148): mean 1.50, 18.2% $\geq$ 2.0. Memory-system refusal, no recitation (n = 240): mean 1.47, 17.1% $\geq$ 2.0. Memory-system substantive engagement (n = 7,835): mean 2.32, 67.2% $\geq$ 2.0. Comparisons: mem-refuse + recite vs. pure no-context refuse, Δ +0.234 [+0.113, +0.355] $p = 0.0002$; mem-refuse no recite vs. pure no-context refuse, Δ +0.206 [+0.103, +0.310] $p = 0.0001$; mem-refuse + recite vs. mem-refuse no recite, Δ +0.027 [−0.098, +0.153] $p = 0.67$.

[90] Tier 2 per-subject per-model responses at `results/_tier2/global_<subject>/`. 5-judge vs 7-judge sensitivity recompute at `docs/research/recompute_5judge_primary.md`. Tier 2 panel-completeness audit (including the 24 GPT-5.4 `FULL_FAIL` cells that drive the 4-judge effective panel in §4.6.1) at `docs/research/v11_panel_completeness_audit.csv`. Mechanical recompute

4.7 Summary of §4 and bridge to discussion

§4 established seven findings:

- **The gradient (§4.1, §4.1.2).** The lower the model's pretraining baseline on a subject, the larger the Spec's lift; the lift in raw points is largest where the baseline is lowest. The gradient holds at the high-baseline end through the Franklin reference.
- **Per-question categorical change (§4.1.1, §4.2.1, §4.4.3).** The Spec produces categorical change on interpretation-required questions and adds little or hurts on literal-recall and refusal-triggering questions. The aggregate effects are conservative averages over batteries that mix the two question types.
- **Compression (§4.2).** The Spec recovers 76% of the corpus's predictive benefit at 4% of the context cost. The structured representation compresses the predictive signal at a fraction of the source-corpus footprint.
- **Content specificity (§4.3).** The effect is content-specific rather than structural. Wrong-Spec adversarial pairings drop accuracy below the no-context baseline ($\Delta = -0.25$); random derangements barely improve on no context (+0.15); only the correct Spec produces the full lift.
- **Memory-system interaction (§4.4).** The Spec layers cleanly on top of three of four commercial memory systems through three patterns (interpretive supply, over-theorization, Spec-induced refusal), with the balance shifting by retrieval architecture.
- **Hedging elimination (§4.3).** Correct-Spec conditions eliminate baseline hedging from 41.2% to 0.4% under the broader-pattern classifier. The hedging-elimination is content-specific, not structure-specific (wrong-Spec preserves baseline refusal patterns).
- **Retrieval divergence (§4.4.1, surfaced post-hoc).** Given an identical fact pool, any two memory systems share zero facts in their top-10 on 35.9% of instances. Mean pairwise overlap is 8.3%. Providers do not converge on which facts are most relevant; recall benchmarks measure recall, while representational accuracy operates at the interpretation layer.

§4.5 reports an exploratory case study on a fundamentally different memory architecture (Letta self-edited block, N=3 subjects); we do not treat the result as a replication or headline finding.

The findings were checked for apparatus artifacts across seven sensitivity tests:

- **Cross-provider response generation (§4.6.1).** Direction reproduces on 7 of 9 (subject, response-model) cells across three response models from two providers (Anthropic Haiku 4.5 + Sonnet 4.6, Google Gemini 2.5 Pro), on batteries regenerated by OpenAI GPT-5.4. Coverage is 3 subjects (Ebers, Yung Wing, Zitkala-Ša), not the full main-study set.
- **Judge panel (§4.6.2).** No directional claim flips between the 5-judge primary and the 7-judge sensitivity, and at the per-subject level no subject's Spec-lift sign changes between panels.
- **Battery composition (§4.6.3).** Neither battery-question-type composition nor Hamerton's leverage explains away the §4.1 gradient slope.
- **Statistical-rigor checks on the gradient (§4.6.4).** Bootstrap CI [−1.25, −0.74] excludes zero and a 50%-attenuated effect. Joint regression (drop Hamerton + literal-recall covariate)

and per-cell panel-range scripts at `scripts/_v10_verification/tier2_mechanical_recompute.py` and `scripts/_v10_verification/tier2_panel_ranges.py`.

holds the slope at −0.87. Permutation test places the observed slope outside the entire null distribution ($p < 0.0001$).

- **Wrong-Spec derangement protocol (§4.6.5).** The wrong-Spec result holds across both protocols (adversarial v1 and random v2).
- **Retrieval-overlap sensitivity (§4.6.6).** The §4.4.1 retrieval-divergence finding survives semantic-similarity matching at every threshold tested and at K=5 as well as K=10.
- **Rubric-handling limitations (§4.6.7).** A post-hoc audit identified two limitations (refusal anchor ambiguity and length-score correlation in C5); both bias the C5 baseline upward, so the reported +0.89 mean lift is conservative rather than inflated.
- **What these checks do not address (§4.6.8).** The class-level LLM-as-judge concern remains and is treated in §6.2.

§5 develops what these results imply for AI personalization beyond the specific experiment, and §6 bounds what the experiment cannot establish.

---

5. Discussion

§4 produced the empirical results; this section discusses their implications for representational accuracy and behavioral alignment. The findings establish that an accurate, interpretive representation of how a specific user reasons improves an AI system's ability to act in alignment with that user, and that recall as the primary metric of AI memory does not capture this dimension. The discussion that follows develops the implications of these findings for AI personalization, particularly as AI moves from a tool to an agent acting on a person's behalf.

5.1 Synthesis: what the seven findings together establish

Across 14 historical subjects and five memory-system configurations, the study tested whether a static interpretive layer[91] increases an AI system's representational accuracy of a specific person. This was operationalized via behavioral prediction on held-out autobiographical text scored by a calibrated, baselined five-judge LLM panel. The layer reliably moves the model from refusal or generic guessing to grounded subject-specific responses where the model has insufficient pretraining on the subject (mean lift +0.89 on the 9 low-baseline subjects; 9 of 9 subjects improved; 78.6% of individual questions improve under the matched layer). 55% of low-baseline questions cross at least one rubric anchor upward, and roughly 1 in 5 cross two or more anchors, meaning the model goes from refusal to a grounded subject-specific prediction in qualitative steps rather than incremental score nudges. The matched layer's content does the work, not the structure of the prompt: an adversarial wrong-Spec control actively degrades performance below baseline. On interpretation-heavy questions where retrieved facts alone underdetermine the answer, the layer supplies the interpretive pattern existing memory systems cannot; three of four commercial systems show positive aggregate prediction-accuracy lift under at least one configuration as a result. The layer recovers most of the predictive accuracy of the full source corpus at 5x to 80x smaller context, and it eliminates response hedging on questions retrieval alone could not ground (41.2% baseline hedging drops to 0.4%). Current memory-system providers do not converge on which facts are most relevant given identical input, even under relaxed similarity matching.

[91] The interpretive layer this paper introduces is the Behavioral Specification: a static document of approximately 7,000 tokens that captures how a specific person reasons. Full pipeline and operational detail in §3.7.

Together, these findings establish that a portable, content-specific, structurally compressible interpretive layer adds a measurable dimension to AI personalization. The layer is distinct from raw facts, raw corpus, current memory-system retrieval, and the pattern-based inferences a frontier model attempts on those inputs on its own; it complements rather than competes with each of them, and makes explicit what the model would otherwise assume implicitly about a specific user. It is most useful where pretraining is thin, but it adds value on top of all three other context types as well, at a context cost compatible with production deployment.

The construct (representational accuracy) has been validated directionally by the data but not absolutely by human annotation; that human-validation follow-up is the highest-priority next step (§7). Robustness checks against cross-provider response models, judge-panel composition, and protocol choices are in §4.6.

5.2 Why the gradient is the load-bearing finding

Almost every living user of AI now and into the future fits the low-baseline band. The population of relevance for AI personalization is the long tail of users whose private reasoning is not in any training corpus. Almost no one in that population looks like Benjamin Franklin; almost everyone looks like the 9 low-baseline subjects, where the layer has the most room to add value. This is the equity property of the approach: the layer brings every user toward consistent representational accuracy, regardless of how thoroughly the pretrained model already knew them.

The implication is straightforward: the less an AI knows about a specific user, the worse it can align its behavior with that user. The gradient shows that providing an interpretive representation moves the AI toward behavioral alignment.[92]

The per-question mechanism beneath the gradient is that the Spec categorically lifts interpretation-required questions while leaving questions the model already answers correctly largely unchanged. This conditional benefit is why the aggregate is not inflated by Spec-induced noise on questions where retrieval already suffices, and why the gradient's slope is real rather than artifactual.

The question that follows is how to ensure the representation is accurate. An individual encounters an effectively infinite set of situations and categories in which they might specify different behaviors or alignments. The challenge is whether a concise structured artifact can compress someone's behavioral patterns with enough fidelity to apply across situations the underlying data never contained. The remainder of this discussion examines what that fidelity requires.

5.3 Retrieval is not interpretation

Recall and preference storage provide real value. The systems that optimize for these tasks do their job well: the four commercial memory systems we tested perform within a few percentage points of each other on recall benchmarks. Interpretation is a different question, sitting at a layer above them.

[92] The slope of the gradient (~ -0.96 in standard regression terms) is a mathematical consequence of post-Spec scores being roughly constant across baselines, not independent evidence that the layer treats different subjects differently. Treatment-heterogeneity readings were considered and rejected in §4.1 on this basis.

The retrieval-divergence finding establishes the separation empirically. Given the same input, the same systems that converge on recall diverge on which facts they surface as most relevant: any two systems share zero facts in their top-10 on 35.9% of question-pairs, with mean pairwise overlap of 8.3% (§4.4.1). Optimizing for recall does not produce interpretation. A memory provider's ranking is a theory of which facts are most similar to a query, computed without a model of how a person reasons.

Interpretation in this study's context has to do with representational accuracy: how well an AI system represents the patterns that shape how a specific person reasons. Stored facts and observed preferences are surface outputs of those reasoning patterns; the interpretive layer is the implicit understanding of the patterns themselves. Recall, preferences, and interpretation are forms of calibrating an AI toward a specific person at different depths: facts and preferences calibrate to surface outputs, while the interpretive layer calibrates to the patterns that produce those outputs. Behavioral alignment requires the deeper calibration. §5.4 picks up what happens when the interpretive layer is composed with memory-system retrieval.

5.4 Composition with retrieval

The implication of §5.3 is not that retrieval is unnecessary, but that retrieval and interpretation each have a distinct role, and a given question may need one, the other, or both in different proportions. The interpretive layer's calibration signal needs to be conditional. The §4.4.3 three composition patterns[93] argue against a static activation rule: always supplying the layer over-theorizes when retrieval already answers and produces rubric-penalized refusals when evidence is insufficient; never supplying it misses the questions where retrieval cannot ground the answer. The three statistical signatures in §4.2.2 sharpen this empirically: adding the Spec to a no-context baseline produces re-ranking (different questions become the well-answered ones, Spearman $\rho = 0.27$), while adding it on top of retrieved facts produces near-uniform lift ($\rho = 0.72$). The Spec does work retrieval does not, and on a different set of questions.

What composition demands, more pointedly, is an understanding of how retrieval and interpretation interact for a specific person on a specific question. Foundational work in the cognitive sciences has long studied how memory and interpretation compose in human reasoning, but this has not been applied to human-AI interaction in any operational way, let alone to scenarios where an AI is acting as an agent on someone's behalf. The interaction dynamics matter most exactly there.

Recent benchmarks reflect a field-level movement toward accurate individual representation. Twin-2K (scaled behavioral prediction), AlpsBench (preference alignment), and PersonaGym (persona fidelity) each measure part of what representational accuracy involves. None of them isolates the interpretation-retrieval interaction directly; that is the gap this paper begins to address.

A candidate architectural step is a serving system that routes between retrieval and interpretation by question type. This implies a framework for categorically distinguishing what kind of information a question actually needs: pure recall (verifiable facts, stored preferences), interpretation (behavioral patterns the person would apply in context), or some mixture. Real-world human complexity does not separate cleanly into these categories; most questions combine them. The architectural

[93] Interpretive supply, over-theorization, and Spec-induced refusal. The Keckley Q21 case in §4.4.4 walks through the third.

answer requires beginning to draw the distinction. §7.4 develops dynamic serving as a production-architecture follow-up.

5.5 Wrong-Spec mechanism and hedging elimination

The wrong-Spec controls in §4.3 establish that the matched layer's content does the work, not the structure of the prompt. Three conditions bracket the finding: a matched layer increases representational accuracy; a random derangement of specifications lands near baseline; an adversarial mismatch (a culturally and temporally distant subject's specification) actively degrades performance below baseline. Structured prompting alone with arbitrary content does not produce the lift; sufficiently mismatched content makes the model worse than no context at all. The §4.6.5 sensitivity check confirms the finding holds across both derangement protocols.

The wrong-Spec adversarial control isolates the content effect from a structural-template reading: under the wrong-Spec, the model often flags the mismatch explicitly rather than complying, and hedging persists at 60.6%. Where the Spec content matches the subject, the model commits; where it does not, the model abstains. Hedging elimination read at the response-style level reflects the same content effect: with the matched layer, baseline hedging drops from 41.2% to 0.4% and the model becomes willing to commit to a specific prediction. The contrast rules out the simplest sycophancy reading: Jain et al. (2025; §2.4) showed that context without the right structure pushes models toward what users appear to want, but our adversarial result actively degrades performance below baseline rather than producing a confident match-something-anything response. The wrong-Spec rules out structural-template sycophancy; the matched layer's content matters.

The Bernal Díaz Q16 case from §4.3 (Example B) shows that some behavioral patterns transfer across people: a wrong-Spec assignment of a culturally distant other's specification can produce a response that aligns with the held-out passage nearly as well as the correct Spec, when the two specifications happen to converge on the same surface prediction by different underlying logics. This convergence at the interpretive layer is informative against §5.3's retrieval-divergence finding. Providers do not converge on which facts are most relevant given identical input; two different specifications CAN converge on the predicted behavior when the underlying behavioral patterns align across people. The two layers operate differently. The Bernal Díaz convergence is direct evidence that the response model is reasoning from the served interpretive content, not pattern-matching to a structural template. What remains unresolved is which structural feature of the layer (anchors, core, predictions, or specific predicates) is doing the work; per-predicate ablation experiments at this scale produced null results consistent with redundant Spec construction, and a controlled component ablation is flagged as the next test in §7.

The matched-layer evidence has implications beyond mechanism. If a representation of this resolution is what produces alignment with how a person reasons, then how it can be compressed to fit production-scale serving (§5.6) and who holds and inspects it (§5.7) become the structural questions.

5.6 Compression and what makes personalization operationally tractable

Compression is what turns the interpretive layer from analytical artifact to deployable system. That a representation roughly 23× smaller than its source corpus recovers most of the predictive accuracy

(§4.2) is the property that makes per-user personalization feasible on current models, where serving the full corpus on every query is not viable. The boundary case where the smallest tested corpus is outperformed by its compressed layer suggests that behaviorally relevant signal is sparse and structurally compressible, not that smaller representations are cheaper to serve.

Faithful compression is the open question. The wrong-Spec controls in §4.3 partially address it: derangements and adversarial mismatches degrade scores below baseline, which is evidence that matched content is doing the work and that arbitrary structure cannot substitute. That is a population-level test, not a per-subject faithfulness measurement. A specification that is small enough to serve and that scores well on a held-out battery has demonstrated compactness and predictive accuracy; structural faithfulness to a specific person's reasoning is a third property, not entailed by the first two. A follow-up is to operationalize faithfulness as its own metric and stress-test compressed representations against the structural patterns that distinguish a person's reasoning.

Compression is a peculiar property to lean on. Regardless of how well a language model uses long context, there is always more context to add; today's models cannot actively serve or construct a representationally accurate understanding of a person from a long-context corpus, which is why compression is load-bearing. That capability gap could close. But even if models acquire that capability, the question is not whether they can use the context well; it is whether they have the right context at all. The conversation shifts from context to representation: the question is which representation the model should reason from, who owns it, and whether it is faithful. Compression makes personalization operationally tractable today; a user-owned, portable, accurate representation of how a specific person interprets and reasons is what the personalization problem reduces to forever.

5.7 Privacy and the case for user-held representation

The inspectability requirement of §1.4 is also an argument about privacy and ownership. Behavioral extraction of the kind this paper performs is a different operation from the data collection that already pervades digital life. Companies record what a person clicks, prefers, and how long they spend on a page. The interpretive layer extracts something deeper: behavioral patterns from raw text, how someone reasons in chat, personal messaging, sustained writing. A representation built from this extraction can produce accurate predictions about a person's future behavior from data that, on its face, looks mundane.

This kind of extraction is already practiced in some contexts. Political and intelligence settings have explored adversarial behavioral profiling for decades; commercial behavioral modeling is incremental but growing. As the operations described in §3 become cheaper and more capable, the asymmetry between what an external party can infer about a person and what that person can see about themselves widens. The institution does not need the representation to be as faithful as the one this paper measures; a crude inference that serves its purpose is sufficient. The findings of this paper say nothing about whether such extraction is ethical; they do say that it is operationally feasible at low cost, with off-the-shelf language models, on a corpus much smaller than what a major commercial platform already holds about its users.

A specific risk frames the inspectability claim from §1.4. A Spec built from publicly available data alone may not capture how a person actually reasons; supplemented with private information, it may misrepresent the person if the public and private data are inconsistent. Either kind of

mismatch can drive the Spec into the adversarial regime established in §4.3. Inspectability and modifiability allow the person to detect and correct these mismatches. They also matter for a deeper reason: a representation that is opaque to the person it represents is a representation that exists, in operational terms, for someone else.

The structural defense against asymmetric behavioral extraction is not to prevent extraction. It is for the person being represented to own the representation. A user-held interpretive layer, inspectable and modifiable, makes the representation visible to the person it concerns. It does not stop external behavioral modeling, but it changes the relationship: a representation produced for the user, owned by the user, and audited by the user exists alongside whatever representations external systems produce. The empirical findings of this paper show that representational accuracy is achievable. Whether it is achieved in service of the person represented or in service of someone else is a structural choice the field has not yet made.

Per-user calibration sits inside the safety envelope, not above or below it. A language model is already shaped toward whatever distribution dominated its pretraining and applies that shape to every user the same way; a Behavioral Specification redirects that existing shape toward the specific user the system is acting on behalf of, rather than adding new shaping on top. Representational accuracy (what this paper measures) and safety alignment (whether a model operates within acceptable bounds regardless of whose instructions it follows) operate at distinct layers, and the constructive question is how the two layers compose. Whether they are formally independent or interact in some structured way is open empirical territory; this paper neither tests nor needs to defend full independence. §7 develops the safety and deployment implications.

5.8 Closing argument

A small authored representation of how a specific person reasons, served as context, changes how a language model predicts that person's behavior on text the model has never seen. The change holds across the 14 subjects measured here, and is largest where the model has the thinnest pretraining signal about the person.

AI memory has been optimized for recall: finding the right fact for a given query. The four leading memory systems, given identical input, do not converge on which facts to surface. Matched content drives the lift, not the structural template of the prompt; arbitrary or adversarial layers degrade the model. Loading facts into context is not knowing someone. The layer this paper measures sits above storage and retrieval: a representation of how a specific person interprets information, distinct from the facts they have produced. Compression makes this layer practical to serve under current context constraints; faithfulness, the question of whether compression preserves the patterns that distinguish a specific person's reasoning, remains a central open question.

The result is not that retrieval is unnecessary, but that retrieval alone leaves a measurable part of memory unmodeled. The narrow claim of this paper is that a compact interpretive representation of a specific person can be authored from source text and improves held-out behavioral prediction beyond what recall over the same source delivers. The paper's claims are directional rather than precise; faithfulness, validation, and generalization beyond this protocol remain open. §7 develops the implications for safety, alignment, and deployment.

---

6. Limitations

The paper's claims are bounded by four axes of constraint on the experimental setup: the subject sample (§6.1), the measurement apparatus (§6.2), the pipeline and specification stability (§6.3), and the scope of exploration (§6.4). Each is a permanent caveat on how the paper's results should be read, distinct from the follow-up experiments proposed in §7.

6.1 Subject sample

The 14 main-study subjects are a selected sample, not a population. Pretraining-coverage bias and the single-living-subject constraint are load-bearing for the paper's framing and are developed in §5.2. This subsection covers four remaining external-validity caveats that §5 does not address.

Public-domain selection. All subjects are historical figures whose autobiographies or memoirs are in the public domain and have been digitized by Project Gutenberg or Internet Archive. That selection pipeline is biased toward figures whose writing was preserved in published form and Western publishing traditions. The paper's cross-continent spread (Saint Augustine, Bābur, Fukuzawa Yukichi, Sunity Devee, Zitkala-Ša, Olaudah Equiano, Mary Seacole) partially mitigates but does not remove this bias.

Self-presentation bias. Autobiographies are self-curation. What each subject chose to include in their memoir is not a neutral record of their behavior; it is a self-selected narrative that may over-represent patterns the author wished to be remembered for and under-represent patterns they chose to leave out. Behavioral-prediction batteries derived from autobiography inherit this bias, and neither the pipeline nor the rubric has a mechanism to correct for it.

Translation artifacts. Three subjects' corpora are English translations of non-English originals (Augustine's *Confessions* from Latin, Bābur's *Bābur-nama* from Chagatai Turkic via Persian, Cellini's autobiography from Italian). Translations introduce stylistic and register shifts that the extraction pipeline processes as if they were original text. A specification authored from a translated corpus may inherit translator choices in addition to the subject's actual patterns.

Era. The oldest subject is 4th to 5th century (Augustine); the newest is early 20th century (Zitkala-Ša, Sunity Devee). Reasoning patterns in modern work contexts, contemporary family structures, technical or digital-native domains, and late-20th-century cultural frames are not sampled. Whether the gradient holds when specifications are authored from modern-era corpora is a generalization the study cannot make from its sample alone.

Autobiography as a genre. Beyond the four caveats above, autobiography is itself a genre with its own distributional properties: reflective, retrospective, narratively shaped, written for an audience, structured by the author's understanding of what makes a coherent life-story. The behavioral patterns extracted from it are patterns *as the author chose to surface them.* Both halves of the train/held-out split come from the same authored narrative, so within-narrative consistency is exactly what the held-out test is positioned to detect. Whether the gradient holds when specifications are authored from non-narrative source material (real-time chat, decisions under time pressure, technical domains the subject never wrote reflectively about) is not testable from this design. The §7.2 living-user replication and §7.5 stateful-agent variants together begin to address cross-genre generalization.

Taken together, these five caveats mean the paper's results should be read as evidence for the claims at the conditions tested. Generalization across era, source language, self-presentation mode, source

genre, and digital-versus-analog source material requires follow-up experiments.

6.2 Measurement apparatus

This section covers the measurement-apparatus constraints on how the paper's numbers should be read. The rubric limitations are in §3.3.6; the LLM-as-judge limitation is the canonical one and is treated in full here.

Class-level LLM-as-judge circularity. Every response in this study is generated by an LLM, every judge is an LLM, and the question batteries are also LLM-generated (§3.5). The 5-judge primary panel and the 7-judge sensitivity check together address within-provider circularity (§4.6.1, §4.6.2): the specification effect reproduces when non-Anthropic response models read non-Anthropic-generated batteries, and removing the most-inflationary judges from the aggregate makes the effect smaller, not larger. What these checks do not address is class-level LLM circularity. The broader concern is that an all-LLM pipeline could be self-reinforcing in ways that human evaluators would not validate. Prior independent work (Zheng et al., 2023) showed that LLM-as-judge panels correlate with human judges on comparable tasks at rates approaching human-human agreement, which is the methodological precedent that legitimizes the panel here. Subsequent panel-based work (Verga et al. (2024) and follow-ons) showed that aggregating multiple LLM judges past a small panel size further tightens agreement. But "approaches human agreement on comparable tasks" is not the same as "is empirically determining the objective quality of a behavioral prediction response." The 5-judge primary panel can answer the directional question (does the specification move representational accuracy in the right direction) but not the absolute-quality question (is any specific numeric value the right score). A stratified human-validation subset is the leading measurement follow-up flagged in §7.1; until that exists, the paper's claims should be read as directional rather than precise. **The paper as a whole is best understood as a methodological prototype with LLM-judge-only evidence on the headline directional claims, awaiting human validation as the highest-priority single follow-up.**

Response-model coverage. The main-study response model is Claude Haiku 4.5. The §4.6.1 Tier 2 cross-provider directional probe ran 2 additional response models (Claude Sonnet 4.6, Google Gemini 2.5 Pro) on 3 subjects spanning the gradient (Ebers, Yung Wing, Zitkala-Ša) against GPT-5.4-regenerated batteries; Claude Opus 4.6, GPT-5.4, and GPT-4o were used as judges in Tier 2 but not as response models. The specification-effect direction reproduced on 5 of 6 (subject, response-model) cells under every panel and Δ-definition tested. The main-study response model is Haiku across all 14 subjects in §4.1; Tier 2 establishes direction across response-model families on a small subset only. The paper's aggregate numbers should be read as what the specification does with Haiku; other response models may produce different absolute magnitudes while preserving the gradient.

Prompt-phrasing ambiguity. The authoring pipeline prompts, the response-generation prompts, and the judge prompts all depend on specific word choices, ordering, and phrasing. We did not systematically test prompt sensitivity as part of this experiment. Prompt sensitivity for the Behavioral Specification authoring pipeline is a separate workstream that informed pipeline design and is distinct from this study's response-generation and judge prompts, which were not varied. Different wordings at any of these stages could produce different numeric results, different extracted fact sets, or different judge scores on the same response. The paper's claims are downstream of the specific prompts used throughout the study (documented in the public repository scripts); we make

no claim about prompt invariance.

Inter-judge calibration variance. Pairwise Spearman ρ across judges is 0.86 to 0.93 (§3.3.4), so the rank order of conditions is stable across the panel. Absolute-score calibration varies (§3.3.3): Gemini Pro fails verbatim-match calibration (4.15 where calibrated judges score 5.0), Opus runs lenient on abstentions (1.41 mean where Sonnet runs strict at 1.14), and length-sensitivity differs across judges. The 5-judge primary aggregate is therefore a stable reading of direction but a panel-specific reading of magnitude. A different judge panel would produce different aggregate numbers while preserving the direction of every claim, which is part of why §5.8 frames the paper as directional rather than precise.

6.3 Pipeline and specification stability

The serving-strategy gap (static full-stack attachment versus production-realistic dynamic activation) is in §5.6 and §7.4. What follows covers pipeline-internal constraints on how the paper's results should be read.

Pipeline version tested. The specifications used in this study were produced by the current pipeline version, which we consider stable. The pipeline has evolved through development, and different pipeline versions produce different specifications on the same source corpus. The paper's results are specific to the pipeline version tested, and the study does not measure how the gradient shifts under earlier or later pipeline versions. The evaluation harness used here can serve as a benchmark for future pipeline iterations: each new authoring-pipeline version can be measured against the current specifications on the same 14-subject batteries to assess whether the gradient strengthens, weakens, or shifts shape (§7).

Specification stability under the same pipeline version. Running the same pipeline twice on the same corpus at temperature 0 does not produce identical specifications. In an initial stability check we ran the full pipeline twice on the same corpus at temperature 0 and compared the two output specifications. Roughly 45% of the resulting text was verbatim-identical between runs. The remaining 55% covered the same predicates and behavioral patterns with different surface phrasing; that 55% was assessed by side-by-side reading rather than against a numeric similarity threshold. This is an artifact of LLM sampling and of the multi-step authoring pipeline: small divergences at the extraction or authoring steps propagate through downstream composition. The per-subject variance probe below replaces this qualitative check with a quantitative one against the rubric.

Per-subject pipeline variance, characterized. A targeted replication probe was run on three subjects spanning the gradient (Sunity Devee, C5 = 1.03; Yung Wing, C5 = 1.88; Augustine, C5 = 2.58). For each subject, the Sonnet layer-authoring step and the Opus compose step were re-run three times against the same per-subject extracted fact set at temperature 0, producing three independent specifications. Each rerun was scored on the full behavioral-prediction battery in the C2a (Spec only) and C4a (facts plus Spec) conditions on the 5-judge primary panel. The resulting per-subject standard deviation of Δ_C4a across reruns is reported below, alongside the cross-subject SD that the §4.1 gradient slope is fit to. For each subject the table reports three numbers: the Spec's effect on representational accuracy in the §4.1 main study (one authored specification), the standard deviation of that effect across three independent pipeline reruns on the same corpus, and that standard deviation as a fraction of the between-subject standard deviation the §4.1 gradient slope is regressed against.

Subject	Canonical Δ_C4a (§4.1)	Per-rerun Δ_C4a SD (n=3)	% of cross-subject SD
Sunity Devee	+1.38	0.103	17.4%
Yung Wing	+0.52	0.055	9.3%
Augustine	+0.11	0.130	22.0%
Pooled (3 subjects)	n/a	**0.101**	**17.1%**

Read of the precision question. The directional finding survives across reruns: low-baseline subjects keep improving (6 of 6 reruns positive across the 2 low-baseline probe subjects), and the gradient slope point estimate is not materially threatened. Quantitatively, the pooled per-subject run-to-run SD of Δ_C4a is 0.10 on the 1-5 rubric, against a cross-subject SD of 0.59 that the gradient slope is regressed against; run-to-run pipeline variance is therefore on the order of 17% of the signal the slope is fit to, and well under the 95% CI half-width on the slope (0.29). What pipeline variance does affect is the precision attached to any single per-subject point estimate. The per-subject Δ_C4a numbers in §4.1 should be read with a soft uncertainty bar of roughly ±0.10 around them. Augustine (mid-baseline, canonical $\Delta = +0.11$) sits at the top of its rerun range and the sign flips on 2 of 3 reruns, so individual mid-baseline subjects' Spec-effect sign is itself within the run-to-run uncertainty band.

Scope and caveats of the variance probe. The probe covers the lighter-scope variance only: the Sonnet authoring step plus the Opus compose step. Extraction-stage non-determinism is held constant by reusing each subject's pre-populated SQLite and ChromaDB state across reruns; including extraction would likely add additional variance at the front of the pipeline. The probe covers low-baseline and mid-baseline subjects but does not reach the Franklin-style high-baseline tail (C5 = 3.77), so the H2 corollary (that the Spec can interfere with strong pretraining signal at the high-baseline end, producing a near-zero or negative Δ) is not directly stress-tested by this run. With n = 3 reruns per subject the per-subject SD point estimates carry their own wide 95% confidence intervals (roughly [0.5×, 6×] of the value); the pooled three-subject estimate is more stable than any single per-subject estimate but should still be read as an order-of-magnitude indicator rather than a precision number. With those caveats stated, the run-to-run SD is small enough relative to the cross-subject SD that we accept the §4.1 slope and R^2 as findings about the gradient rather than artifacts of a single specification authoring.[94]

Pipeline model choices were not varied systematically. The pipeline uses Claude Haiku for extraction, all-MiniLM-L6-v2 for embeddings, Claude Sonnet for layer authoring, and Claude Opus for the compose step (§3.7). These model choices were not varied across the study. Different models at any step could produce different specifications: a different extraction model could surface different facts, a different embedding model could change retrieval behavior, a different authoring model could produce differently-structured anchors and predictions, a different composition model could synthesize the layers differently. Extending model support for each pipeline step, and measuring the gradient under alternate pipeline configurations (for example GPT-5.4 extraction, OpenAI embeddings, a non-Anthropic authoring model), is a direct follow-up flagged in §7.3, alongside the broader question of cross-model consistency for both Spec authoring and usage.

[94] Per-rerun specs and judgments are at `data/global_<subject>/_variance_runs/run_<N>/` and `results/global_<subject>/_variance_runs/run_<N>_*.json`. Full report and reproducibility scripts at `docs/research/v10_pipeline_variance_analysis.md`, `scripts/_v10_pipeline_variance.py`, and `scripts/_v10_pipeline_variance_report.py`.

6.4 Scope of exploration

Not every experimental combination was run. As an independent research project, the study prioritized coverage of the conditions and subjects central to H1 through H5 (§4.1 through §4.4) over running every cell of the design grid. Robustness and ablation conditions were added selectively rather than exhaustively.

Coverage across the experimental grid. The study spans 11 conditions (C1 through C9 plus two wrong-Spec variants), 14 main-study subjects, and a 5-judge primary panel plus 2-judge sensitivity check; response-model coverage is detailed in §6.2 and §4.6.1. Running every possible combination (roughly 6,500 separate cells) was not attempted. Ablation-adjacent conditions (per-layer Spec serving, alternate pipeline model choices, dynamic activation policies) were not run and are planned for future work (§7).

Letta stateful-agent exploration. Letta's stateful-agent architecture is distinct from the archival retrieval path the other three commercial systems use (§4.4, §4.5). Testing the stateful path required a different evaluation harness (§4.5 test design), and that work pulled us partially outside the main-study scope. The resulting comparison covers three subjects (Hamerton, Ebers, Bābur), one Letta version, and one response model (Claude Haiku). Extending the stateful-agent comparison across the full 14-subject gradient, across additional response models, and against future Letta releases is flagged as a follow-up in §7.

Twin-2K is prior work, not a condition of this study. Twin-2K (§2.1) appears in this paper as prior work that measures a related but distinct property (survey-response prediction rather than representational accuracy). We did not run it as a condition of the main behavioral-prediction battery and do not report it as a benchmark result.

---

7. Future Work

Every section of this paper flags at least one follow-up. This section consolidates them into a research agenda organized by theme.

7.1 Measurement methodology

The most impactful measurement follow-up is replacing the content-match rubric with a differentiated battery that separates interpretation-heavy from literal-recall questions and scores epistemic honesty as its own dimension (§3.3.6). Alongside this: a curated question set with explicit quality control on the backward-design process, a human-validated subset of rubric applications to test whether the rubric was reasonably applied per-response (§3.3.6), and human-judge validation on a stratified subset of responses to address class-level LLM-as-judge circularity (§4.6.8, §6.2). Prompt-sensitivity testing across the authoring, response-generation, and judging stages (§6.2) is a separate measurement-stability follow-up that becomes important once the rubric itself is stabilized.

Retrieval-overlap follow-ups (from the surfaced §4.4.1 finding). Two measurement studies remain open after the §4.4.1 sensitivity check that already covers K=5 and semantic-similarity matching for K=10 in both controlled and native configurations:

- **Convergence-at-larger-K analysis.** This study tested K=10 retrieved facts per question (mean Jaccard 0.083 across systems, §4.4.1) and a K=5 sensitivity check that lowered overlap rather than raising it. The follow-up is K=25, K=50, K=100, and higher across the same systems and question set, to map the convergence curve and identify the K threshold at which providers begin to agree on which facts are relevant (if anywhere).
- **Meta-analysis of recall benchmarks against retrieval overlap.** Memory systems that score within a few percentage points on LongMemEval, LOCOMO, and similar recall benchmarks retrieve nearly disjoint top-K facts when given identical fact pools and fixed questions (§4.4.1). Recall benchmarks measure recall, which is what they should measure; the question is what additional dimensions matter for downstream representational accuracy. A meta-analysis comparing benchmark scores to retrieval-overlap on the same systems would clarify what "memory recall" actually predicts about how each system ranks facts for a specific interpretive task. The wrong-Spec per-question meta-analysis (§4.6.5) belongs to the same class of follow-up: a deeper read of which parts of the served context the model referenced under correct versus mismatched specifications.

Confident-misalignment vs abstention. The §3.3 rubric scores both explicit abstention and non-abstention misalignment as 1.00. A coarse post-hoc regex pass classified ~93% of low-end responses as abstention and ~7% as confident-misalignment (wrong referent, off-base inference, or confusion with a different subject; §4.1.1 footnote `[^score-1-composition]`). Why a model picks confident-wrong over abstention on a given question is open: which interpretive frames the wrong-Spec activates that the matched Spec does not, whether the gap correlates with subject-level pretraining coverage, and whether the rubric should score the two failure modes separately are first-order follow-ups.

Pretraining-bleed analysis on wrong-Spec mismatch detection. Under wrong-Spec adversarial pairings, the model often flags the mismatch explicitly rather than complying (60.6% mismatch-detection rate; §4.3). Whether this detection draws on pretraining knowledge of the *correct* subject (effectively recognizing "this Spec doesn't match Lincoln") versus on internal coherence of the Spec content alone is open. A pretraining-bleed analysis would correlate per-subject mismatch-detection rate with per-subject baseline pretraining coverage and with per-pair cultural/temporal distance, isolating which signal is doing the detection. §6.2 flags this as a measurement caveat the present design does not isolate.

7.2 Subject and corpus expansion

A multi-subject living-user replication is the leading follow-up for the entire paper (§5.2, §5.7). The paper's findings are based on 14 historical subjects; whether they generalize to living users is not directly tested by this study, and replicating the gradient with multiple living subjects (with proper consent and privacy infrastructure) would close that gap. Three related expansions: modern-era corpora (to test whether the gradient holds when specifications are authored from contemporary writing rather than pre-20th-century autobiography, §6.1), non-English original sources (to remove translation artifacts, §6.1), and alternative testbeds that isolate reasoning structure without requiring private data, such as U.S. Supreme Court opinions where documented decisions provide a public record of individual interpretive patterns that can be held out and predicted (§5.3).

7.3 Specification design and composition

Component ablation on the authored layers (anchors, core, predictions, brief) is the priority authoring-pipeline follow-up (§5.4). Serving each layer alone and in combinations, measuring Pattern 1 / Pattern 2 / Pattern 3 distributions per configuration, would identify which parts of the pipeline are doing which work. Answers inform both the authoring pipeline's investment priorities and the dynamic-activation policy's weights.

Alongside component ablation: alternate pipeline model choices (extraction, embedding, layer authoring, composition) to measure sensitivity to specific LLM choices at each pipeline step (§6.3); a Base Layer referent-variant that retains named entities inside the same dimensional scaffold, to isolate whether the §4.5 Letta-over-Base-Layer gap is driven by referential vocabulary or by the self-editing process itself (§4.5); and a layered-stack Letta rerun on the matched-rerun subjects, which would likely narrow the §4.5 gap (§4.5).

Cross-model consistency for Spec authoring and usage. Two related questions sit on top of the alternate-pipeline-model-choice follow-up. First, on the authoring side: whether different LLMs at each pipeline step produce specifications that converge on the same behavioral patterns from the same source corpus, or whether the specification itself drifts with the choice of authoring model. Second, on the usage side: whether different response models interpret the same specification consistently, applying it to produce comparable predictions on the same held-out questions. The §4.6.1 Tier 2 probe established that the gradient direction holds across three response models on three subjects; whether the Spec is read and applied the same way across the broader model landscape is the deeper consistency question.

Named-entity grounding as a complement to predicate structure. The §4.5 secondary analysis ruled out surface-syntactic alignment as the mechanism for Letta's case-study lift but identified named-entity grounding as a contributing factor. The Base Layer pipeline abstracts source text into structured predicates for explainability and traceability; specific named entities (place names, person names, institutional names) are deliberately compressed away because they are not behavioral predicates. When held-out questions turn on a particular named entity, this abstraction is a measurable disadvantage. A natural follow-up is a hybrid representation that pairs each predicate with the named entities it grounds in: predicates for structure and explainability, entity tags for referential grounding. The two layers would be indexed to each other so that the response model can pull either or both depending on whether the question turns on the abstract pattern or on the concrete entities the pattern emerged from. A named-entity-grounding-vs-axiom-grounding ablation, plus a paraphrase-resistant rubric that scores directional correctness separately from referential match, are the specific tests this follow-up reduces to.

NLA round-trip Spec faithfulness. Generate hypothetical responses across a battery from a Spec only, then re-extract a Spec from the generated text using the same pipeline. Compare the round-tripped Spec to the original on predicate coverage, axiom recall, and behavioral-pattern fidelity. The round-trip degradation rate measures how much of the Spec's interpretive content survives the model's compression-decompression of it, and whether different model families preserve different aspects (anchors vs predictions, structural vs stylistic) of the same input.

Spec component utilization at runtime. Component ablation tests which Spec components contribute to lift; component utilization tests which components a response model actually grounds in. Activation-level techniques (Anthropic's natural-language-activation work, persona-vector analysis) applied to (Spec, response, ground-truth) tuples could measure whether the model's response

activates the same Spec sections the rubric scores as load-bearing. The measurement bridges representation-design (what the Spec encodes) and response-time use (what the model retrieves and applies).

7.4 Production serving and infrastructure

The study served the Behavioral Specification statically and in full on every query. Four production-realistic serving-layer follow-ups would refine deployment beyond this static-attachment baseline: dynamic activation (selecting which parts of the Spec to attach per query), user editing and inspection (how a user can update or correct their own Spec post-authoring), temporality handling (how the Spec ages and what triggers re-authoring; §7.5), and topic decomposition (whether the Spec can be partitioned by domain for selective serving). Each is a measurement question in its own right: whether the gradient, mechanism, and composition findings hold under each production serving strategy.

Layered-context stacking studies. This paper tests Spec-on-no-context, Spec-on-facts, Spec-on-corpus, and Spec-on-memory-system separately. A stacking study would systematically vary the layered context: memory-system retrieval at increasing K, with and without facts, with and without Spec, with both. The output is a map of which layer combinations produce additive lift, which produce subadditive interaction, and which produce antagonistic effects. The study format extends the §4.4 memory-system layering analysis by treating context type as a factorial design rather than a between-condition comparison.

7.5 Stateful-agent implementations and temporal drift tracking

Several follow-ups sit adjacent to the paper's static-snapshot design.

Stateful-agent variant of the Behavioral Specification. The pipeline as tested is offline and batch. A persistent, self-editing variant that ingests new source material as it arrives, re-edits anchors and predictions in place, and maintains version history with provenance across edits is a natural next step. The §4.5 Letta exploration (N=3, post-hoc) is one data point on an adjacent architecture; building and evaluating a stateful-agent Base Layer implementation on the full 14-subject main-study battery would close the comparison within a single architectural family and extend §4.5 to a layered-stack rerun against Letta at full scope.

Cleaner §4.5 rerun with naming and scaling controls. Two specific extensions of the §4.5 exploration are worth running as a unit. First, anonymize the source corpus before Letta ingestion so Letta writes an anonymized memory block, matching Base Layer's anonymized-during-authoring convention; the §4.5 naming asymmetry (Letta ingests named corpus, Base Layer strips and later restores names) is removed as a confounder. Second, extend the corpus-size axis past the Bābur ceiling to a larger (>250K-word) subject corpus that pushes the Letta block past its character ceiling. The matched-model gap was small at Hamerton (25K-word corpus), largest at Ebers (48K-word corpus), and smaller again at Bābur (223K-word corpus); whether that pattern continues, re-widens, or flattens at extreme corpus size is the empirical question. Both extensions together would turn §4.5's case study into a controlled comparison.

Temporal drift tracking. The static snapshot tested here is a point on a trajectory. A specification authored at one time, compared against a later specification on an expanded corpus, produces a measurable diff: which anchors appear or disappear, which predictive templates shift, which axioms strengthen or weaken. From a sequence of past specifications, the trajectory predicts the

next. The 14-subject corpora collected for this study can be back-sliced by chapter boundaries or publication era for an initial drift test within the current sample. A purpose-built companion study on sequential public records (US Supreme Court opinions, shareholder letters, research papers) is the natural extension.

Canonical life events. Discrete pivots that flip reasoning architecture (a major career change, a religious conversion, a significant loss, a public stance reversal) are distinct from gradual drift. The main-study autobiographies were not structured to test this case. A snapshot specification authored before such an event predicts pre-event reasoning, not post-event reasoning, even though the person's underlying patterns have materially shifted. Whether to detect these events automatically, allow user annotation, or maintain period-specific specifications is an open production-deployment question.

Continuous-representation infrastructure. Both of the above converge on the same engineering target: a background process that watches incoming corpus material, re-authors the specification as new material arrives, and emits drift telemetry as a first-class output (what changed in the Spec, by how much, and against what baseline). The Letta-style stateful agent (§4.5) and the sequential-checkpoint test design are complementary tests for this kind of implementation: the first isolates online self-editing as a way to produce the representation, the second isolates temporality as a property of the representation itself. As a downstream effect, the daemon's continuous output (the sequence of Spec versions, diffs, and drift telemetry) itself becomes a training corpus for next-generation pipeline development.

Additional architectural paths worth testing against the same target, beyond stateful-agent and drift-tracking variants, include agent-edited persistent memories outside the MemGPT family, fine-tuned per-user models that expose their internal representation for audit, and hybrid architectures that combine offline-extracted specifications with online self-editing.

Per-user feedback as a learning signal for the specification. Every interaction with an AI agent can produce a small correction signal: the user edits a response, says "actually I would have done X", or rejects an answer outright. Each of these is a hint that the current specification mispredicted the user's behavior. Used as a feedback signal, these corrections can update the specification directly: the affected predicate or anchor is re-extracted and re-authored, then re-composed into the Spec. Because the update unit is structured text rather than model weights, the change is interpretable and the per-update cost stays at the §3.7 pipeline cost. Two design questions matter. First, drift risk: corrections without a source-text anchor can pull the Spec in arbitrary directions, so the corpus stays the source of truth and corrections are bounded by it. Second, signal quality: explicit corrections are sparse but high-fidelity, while everyday divergence is abundant but noisy; treating both as the same kind of feedback is wrong. A first-pass design using explicit corrections only, batched daily, with versioned diffs logged for rollback, is the cheapest test of whether correction-driven updates produce measurable accuracy gains beyond the static snapshot.

7.6 Safety-alignment integration

The positioning argument (per-user calibration as redirection of existing shaping rather than additional bias, and as orthogonal to safety alignment) is in §5.7. Two concrete follow-ups extend that positioning. First, the Spec-induced refusal cases (§4.3, §4.6.7): a post-hoc classifier audit of 81 Spec-induced refusals across 5 memory systems (`docs/research/refusal_intent_classification.md`) found 75 of 81 (93%) were routine behavioral prediction rather than morally loaded. The refusal pattern is general caution when information is thin, not a moral-integrity mechanism. Whether

it composes cleanly with existing safety frameworks across benign and malicious user types is open. Second, the specifications in this paper were authored from public-domain autobiographies of subjects not selected on intent. What a specification for a user with malicious intent would contain, and what happens when an agent is deployed on that user's behalf, is untested. Both belong to collaboration with AI safety researchers rather than single-lab follow-ups.

A narrower follow-up: the wrong-Spec adversarial control (§4.3) showed response models can recognize Spec content as belonging to a different historical period or persona. Whether models can recognize specifications encoding adversarial values (personas that endorse harmful behaviors) and flag them rather than comply is a direct extension that bears on live-user deployment.

Paper body complete. Abstract to be written last.

---

8. Data, code, and reproducibility

Data availability. All raw response files, per-judge judgments, batteries, and aggregated results for the 14 main-study subjects are in the public study repository at `github.com/agulaya24/beyond-recall` under `results/global_<subject>/` and `results/hamerton/`. Source autobiographies are public domain (Project Gutenberg and Internet Archive). Per-subject Project Gutenberg IDs are listed in §3.4 Table 3.2. Memory-system raw retrieval and ingestion logs are at `results/global_<subject>/<system>_*.json`. The Letta stateful-agent matched-rerun artifacts are at `docs/research/_letta_rerun/`. The full-stack Letta rerun comparison is at `docs/research/_letta_rerun/fullstack_named/`.

Code availability. The Base Layer pipeline source (extract, embed, author, compose) is at `github.com/agulaya24/BaseLayer`. The study-specific analysis and re-run scripts are at `github.com/agulaya24/beyond-recall` under `scripts/`. Reproducibility pointers from each numerical claim to its supporting script are in `docs/PROVENANCE_INDEX.md` and `docs/DATA_REFERENCE.md`. The §4.1 battery-composition sensitivity analysis is reproducible via `scripts/_v10_battery_sensitivity.py`. The §3.3.6 rubric-handling validity audit is reproducible via `scripts/audit_low_end_inflation.py`. The §4.3 hedging classifier is at `scripts/classify_hedging.py`.

Supplementary materials. Per-subject worked examples (Appendix E in the public-repository version) and per-subject paired-delta tables (Appendix B subsections B.2 and B.3 per-subject distribution) live at docs/supplementary/ in the repository. They are cross-referenced from main-paper sections at the indicated pointers and are reproducible from the same data and scripts that produced the in-paper analyses.

Agent-friendly study repo tooling. The study repository is structured for both human reading and agent consumption. A combined SQLite + ChromaDB knowledge index (`workspace/study_knowledge.db`, `workspace/study_vectors/`; built by `scripts/index_study_repo.py`) covers 206 files and 3,702 chunks across the paper, supporting docs, per-subject specs, judgments, retrieval logs, and analysis scripts. An MCP server exposing this index plus typed lookups (per-subject score retrieval, claim provenance, condition-pair anchor-crossing queries) is available as an MVP at `memory-study-repo/mcp/`; the decision to optimize the repo for agent consumption is based on the observation that the same artifacts that make a paper easy to verify mechanically

(stable anchors, structured judgments, machine-readable schemas) also make the paper easier for human readers to skim and verify.

Compute and cost. All response generation and judging used commercial APIs (Anthropic, OpenAI, Google) at standard rates. No specialized hardware was used. All experiments are runnable on a standard developer laptop.

Author affiliation. Aarik Gulaya, Base Layer. Contact: `aarik@base-layer.ai`. Project page: `base-layer.ai`.

Funding. This work was self-funded.

Conflicts of interest. The author is the founder of Base Layer, the project that develops the Behavioral Specification pipeline this paper evaluates. Memory-system providers tested in this paper (Mem0, Letta, Supermemory, Zep) were used through their public APIs at standard rates; no provider was given preferential framing, and Base Layer does not have commercial relationships with any of them. Self-reported benchmark scores from each provider are reported as published; this paper does not adjudicate disputes between providers' published claims (§2.2).

Acknowledgments. Conversations with the broader memory-systems and AI-personalization research communities informed the design of this paper. Specific gratitude goes to the cross-LLM reviewer panels (Gemini 2.5 Pro, Mistral Large, Cerebras Qwen3 235B, Groq Llama 3.3 70B, GPT-5.5) whose iterated reviews materially improved earlier drafts. All errors are the author's.

---

9. References

References

F. C. Bartlett. *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press, 1932.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025.

Prateek Chhikara et al. Mem0: Building production-ready AI agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. arXiv:1503.02531.

Sahaj Jain et al. Interaction context often increases sycophancy in LLMs. *arXiv preprint arXiv:2509.12517*, 2025.

Bowen Jiang et al. Know me, respond to me: Benchmarking LLMs for dynamic user profiling and personalized responses at scale. In *Conference on Language Modeling (COLM) 2025*, 2025. arXiv:2504.14225.

Chris Lu et al. The assistant axis: Situating and stabilizing the default persona of language models. *arXiv preprint arXiv:2601.10387*, 2026.

Adyasha Maharana et al. Evaluating very long-term conversational memory of LLM agents. In *Annual Meeting of the Association for Computational Linguistics (ACL) 2024*, 2024. arXiv:2402.17753.

Charles Packer et al. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Ethan Perez et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022.

Preston Rasmussen et al. Zep: A temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*, 2025.

Vinay Samuel et al. PersonaGym: Evaluating persona agents and LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025. arXiv:2407.18416.

Mrinank Sharma et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

Olivier Toubia et al. Twin-2K-500: A dataset for building digital twins of over 2,000 people based on their answers to over 500 questions. *arXiv preprint arXiv:2505.17479*, 2025.

Pat Verga et al. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024.

Di Wu et al. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations (ICLR) 2025*, 2025. arXiv:2410.10813.

Jingxuan Xiao et al. AlpsBench: An LLM personalization benchmark for real-dialogue memorization and preference alignment. *arXiv preprint arXiv:2603.26680*, 2026.

Lianmin Zheng et al. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS) 2023, Datasets and Benchmarks Track*, 2023. arXiv:2306.05685.

Appendix A. Predicate Vocabulary

A.1 The 46 Constrained Predicates

The extraction step (Step 2 of the pipeline, §3.7) instructs the extraction model to emit triples of the form `(subject, predicate, object)` using only the 46 predicates listed below. Predicates outside this list are either normalized by `normalize_predicate()` into the canonical form or routed to the `unknown` catch-all (which is filterable downstream, not silently dropped). The vocabulary is frozen for the study; it was curated and validated across roughly 50 pilot subjects before being locked. The canonical source is `memory_system/src/baselayer/config.py` lines 613-639 (`CONSTRAINED_PREDICATES`).

The predicates group into seven behavioral dimensions. The groupings below are analytical; the predicate list itself is flat in code.

Behavioral patterns (activities and engagement). These are the most load-bearing predicates for interpretive representation. They describe what the subject repeatedly does or refuses to do, which is what anchors the authored layers in §3.7.

Predicate	Definition	Example usage
`practices`	Repeated deliberate activity, skill-building, or routine.	(subject) `practices` daily writing
`avoids`	Consistent pattern of not engaging with a thing or situation.	(subject) `avoids` hierarchical social settings
`prefers`	Systematic choice of one option over another when both available.	(subject) `prefers` solitary work over committees
`follows`	Active tracking of a person, domain, or source.	(subject) `follows` developments in French art theory
`monitors`	Active observation, narrower than `follows`.	(subject) `monitors` his guardian's health
`plays`	Games, sports, musical instruments.	(subject) `plays` chess
`studies`	Deliberate intellectual engagement with a topic or body of work.	(subject) `studies` Renaissance painters
`builds`	Creation of things, relationships, or institutions.	(subject) `builds` a private library over decades
`manages`	Ongoing oversight or administration.	(subject) `manages` the household finances

Values, beliefs, and self-view. These populate the core layer (§3.7) and describe the stable commitments a subject carries across situations.

Predicate	Definition	Example usage
`values`	What the subject holds as important or worthy.	(subject) `values` intellectual honesty over social standing
`believes`	Propositional commitment, often theological or ideological.	(subject) `believes` scripture is not divinely infallible
`prioritizes`	Revealed-preference ranking under constraint.	(subject) `prioritizes` proximity to family over career advancement
`identifies_as`	How the subject labels or categorizes the self.	(subject) `identifies_as` an independent artist, not a teacher
`aspires_to`	Directional aspiration toward a goal or state.	(subject) `aspires_to` mastery of French prose
`wants_to`	Narrower than `aspires_to`; immediate desire.	(subject) `wants_to` visit the ancestral homes

Emotions and dispositions. These describe affective responses, which often provide the clearest behavioral signal in autobiographical text.

Predicate	Definition	Example usage
`fears`	Things, situations, or outcomes the subject avoids or guards against.	(subject) `fears` religious hypocrisy more than social ostracism
`loves`	Strong positive emotion, stronger than `enjoys`.	(subject) `loves` the moors near Burnley
`hates`	Strong negative emotion, stronger than `dislikes`.	(subject) `hates` formal balls
`enjoys`	Mild positive engagement.	(subject) `enjoys` long walks
`dislikes`	Mild negative response.	(subject) `dislikes` urban environments
`admires`	Respect or admiration, often toward a specific person.	(subject) `admires` Ruskin's early prose
`struggles_with`	Recurring difficulty or area of known weakness.	(subject) `struggles_with` time management
`excels_at`	Recurring demonstrated strength.	(subject) `excels_at` verbal persuasion in small groups

Experiences, decisions, and learning. Transitive episodic events and the inferences drawn from them.

Predicate	Definition	Example usage
`experienced`	An episodic event the subject underwent.	(subject) `experienced` his guardian's death in 1857
`learned`	A concrete skill, fact, or lesson derived from experience.	(subject) `learned` that shame is more effective than instruction in driving mastery
`decided`	A specific documented decision or resolution.	(subject) `decided` not to pursue a political career
`lost`	A thing, relationship, or role no longer held.	(subject) `lost` his fortune in the poetry-book failure
`founded`	Institutions or groups the subject created.	(subject) `founded` a French-language art journal

Relationships (Session 55 expansion). A targeted set of relationship predicates, added to raise relationship-fact extraction from 0.8% to the 3 to 5% range.

Predicate	Definition	Example usage
`married_to`	Marriage relationship.	(subject) `married_to` Eugenie Gindriez
`parents`	The subject's parents (subject is the child).	(subject) `parents` are John and Mary Hamerton

Predicate	Definition	Example usage
`raised_by`	Parental or guardian relationship from the child's perspective.	(subject) `raised_by` his aunt after his father's death
`mentored_by`	Directional mentor relationship; subject was mentored.	(subject) `mentored_by` his guardian's circle
`friends_with`	Friendship.	(subject) `friends_with` a Doncaster schoolfellow
`collaborates_with`	Professional or creative collaboration.	(subject) `collaborates_with` the editor of the *Saturday Review*
`reports_to`	Organizational hierarchy.	(subject) `reports_to` the regimental commander
`relates_to`	Generic relationship fallback when the specific type is unclear.	(subject) `relates_to` the Breadalbane family as distant hosts
`conflicts_with`	Recurring tension or disagreement.	(subject) `conflicts_with` the Anglican social consensus
`maintains`	Ongoing relationship, practice, or commitment.	(subject) `maintains` correspondence with French friends

Biographical context. Stable factual biographical attributes. These are not the most predictive class but are needed for disambiguation and for the anchors layer's detection conditions.

Predicate	Definition	Example usage
`owns`	Property or possessions.	(subject) `owns` a house on Loch Awe
`works_at`	Current or past workplace.	(subject) `works_at` the *Portfolio* magazine
`lives_in`	Current or past residence.	(subject) `lives_in` a Scottish island
`raised_in`	Place where the subject grew up.	(subject) `raised_in` Lancashire
`attended`	Attendance at an institution (may or may not include graduation).	(subject) `attended` Doncaster Grammar School
`graduated_from`	Specifically graduated (distinct from `attended`).	(subject) `graduated_from` no university
`interested_in`	Passive interest, weaker than `follows` or `studies`.	(subject) `interested_in` heraldry

Fallback.

Predicate	Definition	Example usage
`unknown`	Catch-all for extracted claims that do not map cleanly to any of the 45 above. Filterable, never silently dropped.	(subject) `unknown` [unmapped extracted claim]

A.2 Provenance and design choices

The vocabulary was iterated in three stages. The initial 30-predicate list (sessions 1-48) favored values, activities, and biography. Session 49 added 6 predicates (`unknown`, `attended`, `interested_in`, `wants_to`, `loves`, `hates`) to preserve semantic distinctions that the initial vocabulary collapsed. Session 52 added `plays` and `monitors`. Session 55 added 8 relationship predicates to raise relationship-fact extraction from 0.8% to the 3 to 5% target range.

The vocabulary is deliberately behavioral rather than biographical. The ratio of predicates in the behavioral-patterns plus values-beliefs plus emotions-dispositions groups (23 of 46) to the biographical-context group (7 of 46) encodes a design decision: extraction is steered away from facts that are easily verifiable in external sources (city of birth, schools attended) and toward patterns that require reading the source text to infer (what the subject avoids, prefers, values, fears).

A.3 Not in the vocabulary

Three predicate categories that commonly appear in general-purpose knowledge graphs are deliberately excluded:

- **Evaluative predicates about the subject from a third party** (for example, `considered_brilliant_by`). These invert the direction of claim: the subject is the object rather than the source of the reasoning.
- **Time-indexed state changes** (for example, `became`). The vocabulary handles change-over-time through the AUDN ADD / UPDATE / DELETE / NOOP operations at the fact level, not through predicate selection.
- **Causal predicates** (for example, `caused`, `triggered`). Causal inference is produced in the authored layers (predictions, anchors) from collections of facts, not encoded at the extraction step.

A.4 Live deployment

A live web deployment of the pipeline described in §3.7, with served briefs across additional subjects beyond the 14 in this study, is available at base-layer.ai for readers interested in seeing the served-specification format in interactive form.

Appendix B. Question Batteries

B.1 The 10 fixed behavioral-prediction categories

Every behavioral-prediction question in the study is tagged with exactly one of ten fixed categories. The category set is identical across all 15 batteries (14 main-study plus Franklin). Each category is

a behavioral dimension the question probes; the category does not constrain the answer format.

Category	What it probes	Example question
`decisions`	How the subject resolves concrete choices.	"When his uncle's family emigrates to New Zealand, would Hamerton consider joining them?"
`values`	What the subject holds as important when forced to rank.	"When confronted with German Neology, would Hamerton accept or reject the established Protestant position?"
`relationships`	How the subject engages with specific people or classes of people.	"How would Hamerton's religious heterodoxy affect his social standing among the Lancashire gentry?"
`conflict`	How the subject responds when a line is crossed.	"If his tutor attempted to physically harass him, would Hamerton submit or resist?"
`learning`	How the subject acquires skills, knowledge, or lessons.	"Given Hamerton's difficulty following spoken French, what would he do about it?"
`risk`	How the subject handles uncertainty, exposure, or irreversibility.	"Would Hamerton choose to encamp alone on remote Scottish moors, despite it being considered eccentric?"
`creativity`	How the subject produces or evaluates creative work.	"Would Hamerton publish his early poetry at his own expense, and what would the commercial result be?"
`stress`	How the subject responds to pressure, exposure, or failure.	"When offered a grand opportunity to organize an exhibition, would Hamerton accept?"
`career`	How the subject makes professional trajectory choices.	"Would Hamerton follow Ruskin's advice to study nature directly rather than learn from traditional masters?"
`change_over_time`	How the subject shifts or persists across life phases.	"After his poetry failed commercially, would Hamerton continue writing verse?"

Each subject's battery covers 8 to 10 of these categories; no battery skips more than 2. The distribution within a subject reflects what the training half of that subject's corpus naturally supported, not a quota.

B.2 Per-subject battery composition (10-category by 15-subject matrix)

Per-subject battery-composition table (15 subject rows by 11 columns: ten behavioral-prediction categories plus a per-subject total) moved to `docs/supplementary/appendix_B_per_subject_paired_delta_tables.` for length reasons; the headline composition facts are summarized here. Each subject's battery covers 8 to 10 of the ten categories listed in §B.1; per-subject totals are 39 questions for the 13 global subjects and Hamerton, and 40 for Franklin. Column totals across all 15 subjects: decisions 93, values 115, relationships 84, conflict 66, learning 65, risk 19, creativity 34, stress 60, career 23, change_over_time 27, total 586. Raw batteries are at `results/global_<subject>/battery_v2.json` (global subjects), `data/hamerton/battery.json` (Hamerton), and `data/franklin/battery.json` (Franklin); counted over `tier == "behavioral_prediction"` slices.

B.3 Behavioral-axis distribution (LITERAL / INTERPRETIVE / REFUSAL-TRIGGERING)

A secondary classification of the same 586 questions, produced by Claude Haiku 4.5 as a post-hoc auditor, tags each question by what cognitive operation it requires the response model to perform. Full audit at `docs/research/question_category_audit.md`.

Aggregate distribution:

Axis	n	%
LITERAL_RECALL	60	10.2%
INTERPRETIVE_INFERENCE	403	68.8%
REFUSAL_TRIGGERING	123	21.0%

Per-subject distribution: the 15-row per-subject behavioral-axis breakdown (LITERAL / INTERPRETIVE / REFUSAL counts per subject) is in `docs/supplementary/appendix_B_per_subject_paired_delta_` for length reasons; the aggregate distribution above is the load-bearing summary for §4 claims. Notable points from the per-subject table: Hamerton's battery is the most refusal-heavy of the 14 main-study subjects (19 of 39); Franklin's is the most interpretive (37 of 40, with 0 LITERAL_RECALL); the 13 global subjects cluster in the 22 to 33 INTERPRETIVE range with refusal counts between 2 and 15. Source: `docs/research/question_category_audit.md`.

B.4 Category-level effect size on Δ_spec

Mean Δ_spec (C2a minus C5) broken down by the behavioral-axis classification. Source: `docs/research/question_category_audit.md`.

Axis	n	Mean Δ_spec	Median Δ_spec
LITERAL_RECALL	60	+0.792	+0.800
INTERPRETIVE_INFERENCE	366	+0.397	+0.400
REFUSAL_TRIGGERING	120	+0.417	+0.200

The LITERAL_RECALL bucket is small (n = 60) per subject, so per-subject estimates are high-variance, but the aggregate is robust within that constraint. The unexpected finding (that LITERAL_RECALL Δ_spec exceeds INTERPRETIVE_INFERENCE Δ_spec) is discussed in the audit doc: a plausible mechanism is stylistic-register match between the Spec's Victorian prose and Hamerton's held-out text, not genuine fact recall. The INTERPRETIVE_INFERENCE signal (n = 366, Δ = +0.397) is the cleanest between-condition evidence that the Spec performs representational work rather than register-matching.

B.5 Per-subject by axis Δ_spec

Full breakdown at `docs/research/question_category_audit.md` and per-subject axis-Δ scaffold values at `docs/research/v11_emit/appendix_b_battery.json` (claim ids `appB_5_<subject>_<axis>_delta`). Summary of the cross-subject pattern under 5-judge primary aggregation: the strongest positive Spec effects cluster on three subjects. Hamerton (LITERAL +1.68, INTERP +1.30, REFUSAL +1.25), Sunity Devee (+1.38 / +1.16 / +1.35), and Bernal Díaz (+2.00 / +0.44 / +0.64) carry the largest gains. Augustine, Equiano, and Zitkala-Ša show negative or near-zero deltas across all three axes, consistent with their status as mid-baseline subjects on the §4.1 gradient. Fukuzawa and Seacole show their largest positive effects on INTERPRETIVE_INFERENCE specifically (+0.83 and +0.79). Bernal Díaz's +2.00 on LITERAL_RECALL is computed over a small per-axis n (single-digit questions per axis per subject), so the per-subject estimate is high-variance; the cross-subject correlation pattern in §B.6 is the more robust signal.

B.6 Battery-composition sensitivity

This appendix provides the technical detail behind the §4.1 battery-sensitivity controls.

B.6.1 Battery-question-type correlations.

Across the 14 main-study subjects:

- Δ_spec range: −0.31 to +1.37
- Corr(fraction of LITERAL_RECALL questions, subject-level Δ_spec): r = +0.595 (recomputed under strict 5-judge primary; the legacy audit-doc value was +0.646, which used a Hamerton-divergent intermediate aggregation)
- Corr(fraction of INTERPRETIVE_INFERENCE questions, subject-level Δ_spec): r = −0.466 (legacy: −0.582, same caveat)
- Corr(fraction of REFUSAL_TRIGGERING questions, subject-level Δ_spec): r = +0.212

The positive LITERAL_RECALL correlation and negative INTERPRETIVE correlation imply that subjects whose batteries over-weight literal recall also produce larger measured Δ_spec values.

B.6.2 Multiple regression controlling for LITERAL_RECALL fraction.

A multiple regression of Δ_C4a on both C5 baseline and LITERAL_RECALL fraction across the 14 main-study subjects yields a partial coefficient on baseline of $\mathbf{-0.88}$ **[95% CI −1.13, −0.63],** $\mathbf{p < 10^{-5}}$, attenuated from the univariate −0.96 by about 8%. LITERAL_RECALL fraction enters as a significant partial predictor (β = +2.30 [+0.34, +4.26], p = 0.026), but baseline carries the bulk of the explained variance: 63.6% uniquely attributable to C5, 6.9% uniquely attributable to LITERAL_RECALL fraction. The two predictors are not collinear (Pearson r = −0.28, VIF = 1.08 for both), so the partial coefficients are stable. Adjusted R^2 rises from 0.80 to 0.87 when

LITERAL_RECALL fraction is added; the controls are additive rather than redundant. The gradient on baseline survives; it is not an artifact of battery composition.

B.6.3 Hamerton-leverage subset regression.

Hamerton's 80-question battery predates the global-subject pipeline and uses a slightly different backward-design path (the legacy Haiku 4.5 generator that originally produced Franklin and Hamerton); the 13 global subjects' main-study batteries also use Claude Haiku 4.5 but were regenerated by `run_global_rerun.py` against a uniform prompt template. All 14 main-study batteries share the same generator family. A subset regression dropping Hamerton (N=13 globals) yields a slope of $\mathbf{-0.89}$ **[95% CI −1.18, −0.61], $R^2 = 0.81$, $p < 10^{-4}$**, compared to the full-sample −0.96. The point estimate attenuates by about 7%, and the 95% CIs overlap substantially. The gradient is not Hamerton-driven. A separate GPT-5.4-regenerated battery set (`results/global_<subject>/battery_gpt54.json`) exists for each global as a circularity control; its results are reported in §3.5.1 and §4.6.1, not folded back into the §4.1 gradient itself.

B.6.4 Discussion.

This is a battery-composition confound in the cross-subject gradient; the paper's gradient claim is therefore specifically about mean score movement per subject, not about mean score movement per category. §5.3 and §7 flag a follow-up study with a category-balanced battery as the primary design improvement for future gradient work.

B.6.5 Hamerton-leverage at the per-question grain.

The B.6.3 subset regression checks the per-subject mean grain. A parallel question is whether the per-question extreme upward anchor crossings catalogued in `docs/research/wins_inventory_20260428.json` (60 unique cases across 18 condition pairs, 351 paired low-baseline questions on C5 to C4a) are concentrated on Hamerton specifically. They are: Hamerton accounts for 15 of the 60 unique extreme jumps (25%) on a battery of 80 questions (18.75% extreme-jump rate); the other 13 subjects average 8.9% extreme-jump rate across 39-question batteries. Hamerton's elevation is real but its cause is not isolated by the present design. Three candidate mechanisms (legacy battery-generator path, subject pretraining thinness, behavioral-predicate density per word) are not separately identifiable, since Hamerton's served Spec is 1918 words (brief-only) versus globals' ~5775 words (anchors + core + predictions + brief), so Spec length is anti-correlated with extreme-jump rate. As measured by the heuristic classifier, the mechanism distribution on Hamerton's 15 jumps versus globals' 45 is nearly identical (PATTERN_PREDICATE+HYBRID share: Hamerton 73.3% vs globals 80.0%); the heuristic does not discriminate jumps from non-jumping Spec-loaded controls (`docs/research/pattern_activation_deep_20260428.md`), so this near-identity is consistent with the heuristic detecting Spec-loaded response style rather than the lift mechanism.

B.7 Coupling-free reframing of the gradient

The headline slope regresses Δ_C4a = C4a − C5 on C5, which mechanically embeds a −1 component when C4a is bounded on the 1-5 scale and partially independent of C5. To triangulate from a non-coupling-prone angle, we ran three additional checks on the same per-subject (C5, C4a) data (script: `scripts/_v10_coupling_sensitivity.py`; full output: `docs/research/v10_coupling_sensitivity_analysis.md`).

B.7.1 Level regression.

The level regression C4a ~ C5 produces a slope of $\mathbf{+0.04}$ [95% CI −0.24, +0.33], $R^2 = 0.008$, p =

0.76. C4a is essentially flat across the C5 range of 1.02-2.77 and clusters tightly around its mean of **2.46** at the per-subject grain. The Spec does not differentially "lift" low-baseline subjects more than high-baseline ones in any treatment-effect-heterogeneity sense; it produces a roughly constant post-Spec C4a mean per subject regardless of baseline, and the apparent Δ-on-C5 gradient equals the baseline shortfall.

B.7.2 Permutation test.

A 10,000-iteration permutation test that shuffles C4a across subjects (preserving the bounded marginal but breaking any link to C5) yields a null distribution for the Δ-on-C5 slope centered at −0.998 (SD 0.127). The observed −0.960 is not extreme against this null (two-sided p = 0.77). In plain language: even when C4a values are randomly reshuffled across subjects, the Δ-on-C5 slope still lands near −1 on average, because the change-score parameterization mechanically pushes the slope toward −1 whenever C4a is roughly independent of C5. The −0.96 the headline regression reports is what the regression arithmetic forces, not independent evidence that low-baseline subjects benefit more from the Spec at the per-subject mean grain.

B.7.3 Bootstrap.

A 10,000-iteration subject-level bootstrap returns CIs of [−1.254, −0.740] for the Δ-on-C5 slope and [−0.254, +0.260] for the level slope. The level CI straddles zero, consistent with the level-regression finding that the per-subject C4a mean is roughly constant across baselines.

B.7.4 Reading the gradient against this.

The substantive finding survives the coupling check, but its framing has to shift away from "the Spec acts more strongly on low-baseline subjects" toward the per-question reframing in §4.1: low-baseline subjects have a larger pool of questions at low rubric anchors, so the Spec has more opportunity to produce upward integer-band crossings, which aggregates as a larger per-subject mean lift. The directional asymmetry on those crossings (no observed transitions from bands 2, 3, or 4 into band 5 across the full 14-subject panel; the only band-5 endpoints reached come from band 1) is consistent with the §4.2 finding that even the full source corpus C8 plateaus at a similar per-subject mean.

B.8 Per-predicate ablation (Phase 2c)

To probe whether single behavioral predicates within the Spec are uniquely load-bearing, we ran a per-sentence ablation experiment on a stratified sample of 16 extreme-upward-jump cases. For each case, the heuristically-identified causal predicate (highest-token-overlap Spec sentence vs the question and held-out passage) was located in the served Spec and three response variants were generated at temperature 0 with Claude Haiku 4.5: (1) original (full Spec), (2) ablated (predicate removed, replaced with a length-matched neutral biographical filler), (3) reversed (predicate replaced with a behavioral opposite synthesized by Sonnet). Each variant was scored by the 5-judge primary panel.

Results (script: `scripts/run_predicate_ablation.py`; data: `docs/research/predicate_ablation_results_2`

- Mean Δ_removal (original minus ablated) across 16 cases: +0.05 anchor points (95% CI [−0.35, +0.45])
- Mean Δ_reversal (original minus reversed): −0.24 anchor points (95% CI [−0.45, −0.02])
- 2 of 16 cases showed Δ_removal ≥1 anchor; 11 of 16 had Δ_removal < 0.5

Single-predicate removal does not measurably reduce response quality on this sample. The paper

does not interpret this as evidence that the Spec is mechanistically inert: the higher-level mechanism evidence from the wrong-Spec adversarial control (Appendix C / §4.3) shows the Spec as a whole is doing causal work. The null result on per-sentence ablation is consistent with redundant Spec construction, in which multiple sentences across the anchors / core / predictions / brief layers reinforce the same behavioral patterns; removal of any single sentence leaves the pattern accessible elsewhere in the Spec.

A methodological caveat applies. Original-condition reproduction at temperature 0 was not bit-exact deterministic across reruns; mean drift between the recorded original score (from `docs/research/wins_inventory_20260428.json`) and the rerun original score was −1.44 anchors, with 9 of 16 cases drifting by more than 1 anchor. Some of the variance in Δ_removal is rerun stochasticity rather than ablation effect. The extreme-upward-jump cases specifically show higher pipeline variance than the per-subject mean grain documented in §6.3.

Future work tightening (per the test's own report): human-rated predicate identification (vs heuristic), larger N (all 47 PATTERN_PREDICATE cases), irrelevant-predicate control (matched-length unrelated predicate to test the "any rich persona text" alternative), multi-predicate cluster ablation.

B.9 Footnote-redirect technical detail

This subsection holds the longer technical content for footnotes that would otherwise grow to multi-paragraph length. Each entry is keyed to the footnote name in the body.

B.9.1 `[^delta-aggregation]`. +0.89 vs +0.93 reconciliation.

The +0.89 figure is the canonical cross-subject mean of per-subject Δ_C4a. Each subject's Δ is computed as that subject's per-question 5-judge primary mean under C4a minus their per-question mean under C5; these per-subject Δs are then averaged across the 9 low-baseline subjects. The grand-mean alternative grand-averages all per-question scores under each condition first and then takes the difference, yielding +0.93 (the difference of the C4a grand mean 2.45 and the C5 grand mean 1.52). The two numbers are not in conflict; they answer slightly different questions. The per-subject-mean grain (+0.89) is the unit of inference used throughout this paper because every statistic is computed at the subject level first, then aggregated across the 14 subjects (§1.2 aggregation rule).

B.9.2 `[^heldout-leakage-audit]`. Held-out leakage audit detail.

A held-out leakage audit on the 60 unique extreme-upward-jump cases (full report at `docs/research/held_out_leakage_investigation_20260428.md`) found 0 6-gram, 2 4-gram, and 12 3-gram matches between held-out passages and C4a responses. Of the 9 cases with any leak, 6 are short generic phrases also resident in the served facts list (CORPUS_LEAK), 2 are subject-specific n-grams not in any served context (best explained by pretraining recall of public-domain autobiographies; PRETRAINING_MEMO_CANDIDATE), and 1 is generic English (COMMON_PHRASE). The longest shared run anywhere is 4 tokens, well below transcription length. Severity verdict: rare; no structural validity concern; footnote acknowledgement is the appropriate paper-text treatment. Excluding the 2 pretraining-memorization candidates from the extreme-upward-jump set shifts the C5 to C4a low-baseline extreme-jump count by at most 1 (20 to 19); per-subject mean Δs are unchanged at the per-question level. The "held-out passage" was held out from served Spec / facts, not from pretraining, and the audit confirms that interpretation: where C4a held-out-to-post leakage exists, it is either short generic phrasing also resident in the

served facts (trivially short) or subject-specific content in the model's pretraining, not study-design contamination of the served context.

B.9.3 `[^supermemory-no-retrieval]`. Supermemory NO_RETRIEVAL placeholders.

Across the full 14-subject Supermemory analysis, 30 individual responses (Augustine 2 questions, Equiano 28 questions) were Supermemory provider-failure placeholders rather than substantive predictions, scored at the rubric floor (1) by the judge panel. We treat these as scored data rather than missing data, consistent with how the rest of the study handles low-quality responses. Excluding the 30 NO_RETRIEVAL records as missing data would shift Supermemory's aggregate Δ slightly higher; the qualitative story (small aggregate at both grains, bimodal per-question distribution) holds either way.

B.10 Pre-registered hypotheses and post-hoc analyses

The paper distinguishes pre-registered hypotheses (H1–H5, locked against the statistical commitments in `docs/ANALYSIS_PLAN_LOCK.md`) from analyses that emerged during the work. The following table catalogues every load-bearing analysis result reported in §4 and identifies its status. Post-hoc items are reported as exploratory rather than at the same evidentiary tier as the pre-registered hypotheses.

Item	Status	Where reported	Note
H1 Spec-context outperforms no-context	Pre-registered	§4.1, §1.3 1st bullet	Headline gradient
H2 Spec benefit inversely proportional to pretraining coverage	Pre-registered	§4.1, §4.1.2, §1.3 1st bullet	Gradient at both ends; Franklin reference
H3 Content-specificity (correct vs. wrong Spec)	Pre-registered	§4.3, §1.3 4th bullet	Wrong-Spec controls v1 + v2
H4 Spec interacts with retrieval through three patterns	Pre-registered	§4.4, §4.4.3, §1.3 5th bullet	Memory-system composition
H5 Compression: ~7K-token Spec recovers most of corpus signal	Pre-registered	§4.2, §1.3 3rd bullet	At 5x to 80x smaller context
Cross-system retrieval-overlap divergence	Post-hoc	§4.4.1; sensitivity in §4.6.6; §1.3 7th bullet	Surfaced during memory-system analysis; mean Jaccard 0.083 across 10 system pairs; survives semantic-similarity matching
Letta stateful-agent case study	Post-hoc	§4.5; full in Appendix G	N=3, exploratory
Letta semantic-duplication scaling	Post-hoc	§4.5; Appendix G	Surfaced in this paper's analysis; cosine $\geq 0.85 = 56.1\%$ on Bābur
Abstention-credit validity audit	Post-hoc	§3.3.6	9.4% of refusals score ≥ 2.0; bias direction makes the Spec effect likely larger than reported
Per-subject wrong-Spec heterogeneity	Post-hoc	§4.6.5	5/13 subjects show small positive v1 deltas (coincidental content overlap)
Hedging-elimination (28.8% $\rightarrow$ 0.0%)	Post-hoc	§4.3, §1.3 6th bullet	Surfaced from response-level audit
Battery-question-type sensitivity (literal-recall fraction)	Post-hoc reactive	§4.6.3, Appendix B.6	Added in response to v9/v10 reviewer concerns
Hamerton leverage check (subset regression)	Post-hoc reactive	§4.6.3, Appendix B.6	Added in response to v9/v10 reviewer concerns
Coupling-free reframing of the gradient	Post-hoc reactive	§4.1.1 leveler callout, Appendix B.7	Added in response to GPT-5.5 review

Item	Status	Where reported	Note
Cross-provider response generation (Tier 2)	Pre-registered control	§3.6, §3.5.1, §4.6.1	Sonnet 4.6 + Gemini 2.5 Pro on 3 subjects
GPT-5.4 battery regeneration (Control 1)	Pre-registered control	§3.5.1, §4.6.1	Battery generator circularity
Judge-panel composition (5-judge primary, 7-judge sensitivity)	Pre-registered control	§3.3.3, §4.6.2	Locked panel before scoring
Wrong-Spec derangement protocol sensitivity (v1 vs v2)	Reactive	§4.6.5	v2 is the standard randomization control; v1 is the adversarial stress test (headlined for stronger evidence)

Reproducibility scripts and raw data for each row are pointed to throughout §4 and consolidated in §8 Data, Code, and Reproducibility.

B.11 Per-system per-subject paired-delta distributions

The §4.4.3 footnote `[^memsys-pattern-appendix]` collects the per-cell counts behind the three-pattern claim. This subsection consolidates those cells into a single table and a short reading note. The unit of observation is a single (subject, question) pair scored under both retrieval-only (C1) and retrieval + Behavioral Specification (C3); per-cell counts are restricted to questions with 5-judge primary coverage on both conditions. "Increases" / "decreases" use the $|\Delta| \geq 1.0$ threshold on the 5-point rubric (one full anchor crossing) so that small judge-noise jitter does not inflate the count in either direction.

Memory system	Subject	Aggregate Δ_spec	Increases (Δ ≥ +1.0)	Decreases (Δ ≤ −1.0)	Net at the per-question grain	Source
Supermemory	full 14-subject pool	≈ 0 (closest to zero)	57	53	110 of 546 paired questions cross by ≥ 1.0 (20.1%); the two roughly cancel at the mean	§4.4.3 lede paragraph + [^supermemory-scaffold]
Mem0	Yung Wing	+0.33	21	10	31 of 39 paired questions cross by ≥ 1.0; 21 helps outnumber 10 hurts 2.1×	[^memsys-pattern-appendix]
Mem0	Keckley	−0.02	12	13	25 of 39 paired questions cross by ≥ 1.0; counts even at 12 / 13, aggregate near zero	[^memsys-pattern-appendix]
Letta archival	Hamerton	+0.42	19	7	26 of 39 paired questions cross by ≥ 1.0; 19 helps outnumber 7 hurts 2.7×	[^memsys-pattern-appendix]

Memory system	Subject	Aggregate Δ_spec	Increases (Δ ≥ +1.0)	Decreases (Δ ≤ −1.0)	Net at the per-question grain	Source
Zep	Seacole	+0.47	20	7	27 of 39 paired questions cross by ≥ 1.0; 0 questions show large regressions in this cell	[^memsys-pattern-appendix]
Base Layer	Yung Wing	+0.29	19	7	26 of 39 paired questions cross by ≥ 1.0; 19 helps outnumber 7 hurts 2.7×	[^memsys-pattern-appendix]

Reading. Every cell is a mixture. Strong-positive aggregates (Letta Hamerton, Zep Seacole, Base Layer Yung Wing) still contain 7 large regressions per cell. Near-zero aggregates (Mem0 Keckley) resolve into substantial counts in both directions rather than a flat per-question profile. The Supermemory pool is the cleanest read on the three-pattern mixture because the helps-versus-hurts counts are nearly balanced; the same shape reproduces across every other (system, subject) cell in the study.

The cells above are representative rather than exhaustive (one cell per memory system, two for Mem0 to capture both a positive-aggregate and a near-zero-aggregate subject). Full per-system per-subject paired-delta arrays are at `docs/research/per_system_anchor_crossing_20260427.json`; the recompute script is `scripts/_table_4_6_5judge_recompute.py`.

Appendix C. Conditions, Models, and Memory-System Configurations

C.1 Condition identifiers (summary card)

A consolidated lookup for the condition IDs used throughout §4. Defined in §3.2; summarized here.

ID	Family	Context served	Purpose
C5	Direct	None. Question only.	Pretraining-only floor. Baseline.
C2a	Direct	Behavioral Specification only.	Isolate Spec's contribution.
C2c	Direct	A random other subject's Spec (derangement, seed=42).	Wrong-Spec control.
C4	Direct	Full extracted fact set for subject.	Raw fact volume, no structure.
C4a	Direct	Full facts plus Spec.	Spec added to raw facts.
C8	Direct	Full training corpus (half of source text).	Uncompressed source.
C9	Direct	Training corpus plus Spec.	Spec added to raw source. Bābur excluded (422K word overflow).
C1	Memory system	Top-k retrieval output from system.	Retrieval only, each of 5 systems.
C3	Memory system	Top-k retrieval output plus Spec.	Retrieval plus Spec, each of 5 systems.
C1_<system>_fullpipeline	Memory system, native	Retrieval from system-native ingestion of raw corpus.	Native ingestion variant, retrieval only.

ID	Family	Context served	Purpose
C3_<system>_fullpipeline	Memory system, native	Native ingestion retrieval plus Spec.	Native ingestion variant, retrieval plus Spec.

The `<system>` slot ranges over {mem0, letta, supermemory, zep, baselayer}. Base Layer is run in a single controlled configuration; the four commercial systems are run in both controlled and native variants.

C.2 Shared response-model invocation

Every response call, across every direct-context and memory-system condition, uses the following parameters:

Parameter	Value
`temperature`	0
`max_tokens`	1024
System prompt	Framing instruction: predict how `<subject>` would respond; answer in subject's voice, grounded in demonstrated patterns.
User prompt format	`<context block>\n\nQuestion: <question text>`
Context block	Condition-dependent. Empty in C5. Spec in C2a. Wrong Spec in C2c. Facts in C4. Facts plus Spec in C4a. Corpus in C8. Corpus plus Spec in C9. Retrieval output (optionally plus Spec) in C1 and C3.

No prompt instruction coaches the model to abstain, hedge, or commit. The model's refusal-or-commitment pattern given a specific context is part of the phenomenon being measured (§3.6, §4.3).

C.3 Response models

Role	Model identifier	Provider	Scope
Primary response	`claude-haiku-4-5-20251001`	Anthropic	All 14 subjects, every condition. Main study.
Tier 2 response	`claude-sonnet-4-6`	Anthropic	3 subjects (Ebers, Yung Wing, Zitkala-Ša), C5 / C2a / C2c / C4a against GPT-5.4 batteries.
Tier 2 response	`gemini-2.5-pro`	Google	Same 3 subjects, same conditions as Sonnet Tier 2.

Source: scripts/run_global_subjects.py, scripts/run_full_study.py, scripts/run_multimodel_response

C.4 Pipeline models (specification generation)

Pipeline step	Model identifier	Temperature	Purpose
Extract (Step 2)	claude-haiku-4-5-20251001	0	AUDN fact extraction, 46-predicate constrained vocabulary.
Embed (Step 3)	all-MiniLM-L6-v2 (local)	n/a	ChromaDB vector index (L2 distance).
Author (Step 4)	claude-sonnet-4-6	0	Three authored layers (anchors, core, predictions). Blind regen, domain guard.
Compose (Step 5)	claude-opus-4-6	0	Unified brief composition.
Battery generation	claude-haiku-4-5-20251001	0	Backward-design from held-out corpus.
Battery generation (circularity control)	gpt-5.4 (via OpenAI API)	0	Independent regeneration on 13 global subjects.

Source: memory_system/src/baselayer/config.py.

C.5 Judge panel

Judge	Model identifier	Provider	In 5-judge primary?	Calibration performed?
Haiku	claude-haiku-4-5-20251001	Anthropic	Yes	Yes
Sonnet	claude-sonnet-4-6	Anthropic	Yes	Yes
Opus	claude-opus-4-6	Anthropic	Yes	Yes
GPT-4o	gpt-4o-2024-08-06	OpenAI	Yes	Yes
GPT-5.4	gpt-5.4	OpenAI	Yes	Yes
Gemini Flash	gemini-2.5-flash	Google	No (sensitivity only)	Yes
Gemini Pro	gemini-2.5-pro	Google	No (sensitivity only)	Yes

Judges are invoked independently (no cross-judge conditioning). Each judge sees: held-out ground-truth passage, subject context (name, source), question, response. Judge temperature 0. Judge output is a numeric 1-5 score plus a free-text justification. Calibration diagnostic results in §3.3.3.

C.6 Memory-system ingestion and retrieval parameters

Controlled configuration (C1 / C3) holds the input identical across systems: each system receives the same extracted fact set used by the Base Layer pipeline, re-ingested through its own API. Native configuration (_fullpipeline) has each system ingest the raw training corpus directly via its own

chunking and extraction.

System	Ingestion endpoint	Ingestion unit (controlled)	Ingestion unit (native)	Retrieval top-k	Notable configuration
Mem0	`POST /v1/memories/`	One fact per POST	Raw corpus chunks (Mem0 chunker)	10	`infer=False` on controlled (store as-is, no reformulation). Failure mode: Mem0 may reformulate on `infer=True`, flagged in `docs/PROVIDER_ISSUES.md`
Letta (archival)	`POST /v1/agents/<id>/archival`	One fact per `passage` via `save_memory` tool	Letta native chunking	10	1 fact = 1 passage. Batch ingestion tested 135x faster but changes chunking behavior (see `run_memory_system.py` line 456-458).
Letta (stateful)	Agent state edit during ingestion	One fact per edit cycle	Raw corpus	n/a (read from block)	Evaluated as a separate path in §4.5, not as a row in the C1 / C3 conditions.
Supermemory	`POST /v3/memories`	One fact per memory, `containerTags=<subject>`	Raw corpus	10	`limit=10` on retrieval.
Zep	Graph ingestion via `zep_client.graph.add`	One fact per edge	Raw corpus	10	Retrieval via `client.graph.search(user_id=..., query=..., limit=10)`.

System	Ingestion endpoint	Ingestion unit (controlled)	Ingestion unit (native)	Retrieval top-k	Notable configuration
Base Layer	Direct into ChromaDB	One fact per vector	n/a (Base Layer has no native variant)	10	MiniLM embeddings, L2 distance, cosine-like similarity via `1 - dist^2/2`.

All five systems use the same top-k of 10. All five are queried with the question text as the query. All five feed their retrieval output into the standard prompt schema (§C.2) as the context block.

C.7 Ingestion exclusions and failure cases

Subject / system	Issue	Resolution
Bābur, C9 (raw corpus plus Spec)	422,772-word source exceeds Haiku context window.	Excluded from C9. 13 of 14 subjects report C9 numbers.
Letta native (all subjects)	Ingestion ceiling on archival passages; retrieval produces 0.34-0.47 dedup ratio, meaning a top-10 list often contains 3-5 unique facts.	Reported as-is in §4.4. Not excluded.
Mem0 native	Mem0's `infer=True` reformulated facts during native ingestion pilot.	Used `infer=False` on controlled configuration to hold input identical. Native variant retains `infer=True` (the realistic deployment path).
Zep graph bias	Zep graph retrieval surfaces entity-dense chunks over behavior-dense chunks.	Reported as-is. See `docs/PROVIDER_ISSUES.md`.

C.8 Analysis plan lock

The condition matrix was frozen in `docs/ANALYSIS_PLAN_LOCK.md` before scoring. Any condition added after the lock is reported separately (for example, the Tier 2 3-subject replication and the v2 random-derangement wrong-Spec draws).

Appendix D. Validity Audit and Score Distributions

D.1 Per-subject 5-judge primary aggregate (main gradient)

This table reproduces the §4.1 cross-subject gradient for reference. Every number is the 5-judge primary mean (Haiku, Sonnet, Opus, GPT-4o, GPT-5.4) over the 39-question behavioral-prediction

battery per subject (40 for Franklin).

Subject	Baseline (C5)	Spec only (C2a)	Facts + Spec (C4a)	Δ_spec	Δ_facts+Spec	Anchor crossed
Ebers	1.02	1.54	2.07	+0.52	+1.05	yes
Sunity Devee	1.03	2.27	2.41	+1.24	+1.38	yes
Hamerton	1.26	2.63	2.77	+1.37	+1.51	yes
Fukuzawa	1.67	2.35	2.78	+0.68	+1.11	yes
Bernal Díaz	1.70	2.27	2.48	+0.57	+0.78	partial
Bābur	1.76	1.91	2.01	+0.15	+0.25	no
Seacole	1.77	2.48	2.59	+0.71	+0.82	yes
Keckley	1.84	2.43	2.44	+0.58	+0.59	no
Yung Wing	1.88	2.22	2.40	+0.34	+0.52	no
Zitkala-Ša	2.34	2.03	2.02	−0.31	−0.32	no
Cellini	2.38	2.54	2.53	+0.16	+0.15	no
Rousseau	2.44	2.81	2.53	+0.37	+0.10	no
Augustine	2.58	2.48	2.70	−0.11	+0.11	no
Equiano	2.77	2.46	2.42	−0.31	−0.35	no
Franklin (control)	3.77	3.37	3.65	−0.40	−0.13	no

Raw per-judge files: `results/global_<subject>/judgments_v2.json` and `*_judgments_<judge>.json` (per-judge) for the 13 globals. Hamerton: `results/hamerton/*_judgments_<judge>.json`. Franklin: `results/franklin_legacy_20260411/analysis/*_judgments.json`.

D.2 Per-subject anchor-crossing on the low-baseline slice

Anchor-crossing rate is the fraction of per-question paired (C5, C4a) responses where the C4a 5-judge primary mean lands in a different integer rubric band than the C5 mean. Definition in §3.3.1 and `scripts/compute_anchor_crossing.py`.

Slice-level:

- Total low-baseline questions (9 subjects, 39 Q each): 351
- Upward crossings: 193 (55.0%)
- Downward crossings: 24 (6.8%)
- Stayed in band: 134 (38.2%)

Per-subject breakdown (5-judge primary, paired C5 vs. C4a over N=39 per subject):

Subject	Upward	Upward %	Downward	No crossing
Sunity Devee	29	74.4%	0	10
Hamerton	27	69.2%	0	12
Fukuzawa	26	66.7%	3	10

Subject	Upward	Upward %	Downward	No crossing
Bernal Díaz	23	59.0%	3	13
Seacole	21	53.8%	3	15
Ebers	19	48.7%	0	20
Keckley	19	48.7%	6	14
Yung Wing	19	48.7%	5	15
Bābur	10	25.6%	4	25
Slice total	**193**	**55.0%**	**24**	**134**

Eight of the nine low-baseline subjects cluster in the 48-74% upward band. Bābur is the low-baseline outlier (source corpus 422K words, partial pretraining exposure); he is the only subject whose upward-crossing rate falls below 48%, and his downward-crossing count (4 of 39) is mid-range. Sunity Devee's 74.4% upward rate is consistent with her unusually low C5 baseline of 1.03 noted in §4.1. Per-subject downward-crossing rates stay at or below 15% for every low-baseline subject. Source: `scripts/compute_anchor_crossing.py` executed against `results/global_<subject>/judgments_v2.json` and `results/hamerton/`.

D.3 Rubric-handling validity audit (full report)

This audit is the formal report that §3.3.6 summarizes. It was produced by `scripts/audit_low_end_inflation.py`. Raw flagged cases live in `docs/research/s114_low_end_inflation_audit.json`; source response and judgment data are under `results/global_<subject>/`. The audit is restricted to the 9 low-baseline subjects (1,599 responses across C5, C2a, C2c, C4, C4a conditions).

D.3.1 Abstention detection.

Abstention patterns were matched by regular expression against response text. Pattern list includes variants of "I don't have specific information," "there is no explicit documented," "I cannot confirm," "I am not certain," "would need additional context," "my training data does not," and related phrasings. Full pattern list in `scripts/audit_low_end_inflation.py` lines 29-42. 192 of 1,599 low-baseline responses (12.0%) matched one or more abstention patterns.

D.3.2 Score distribution of abstention-matching responses.

The rubric's lowest anchor is "refuses or off-base." An honest refusal should score at or below 1.5 (closer to rubric-1 than to rubric-2). The distribution of 5-judge primary means over the 192 abstention-matching responses:

5-judge primary band	Count	% of abstentions
1.0 to 1.5	159	82.8%
1.5 to 2.0	15	7.8%
2.0 to 2.5	12	6.3%
2.5 to 3.0	2	1.0%
3.0 to 3.5	2	1.0%
3.5 and above	2	1.0%

82.8% of abstentions score in the expected band. 18 of 192 abstentions (9.4%) score at or above 2.0,

and 6 of 192 (3.2%) score at or above 3.0. Mean abstention score: 1.27 (expected: close to 1.0). Under a clean rubric these would all be closer to 1.0; the over-credit reflects judges giving partial marks for adjacent-fact recitation or for correctly identifying what the context does not contain.

D.3.3 Per-judge strictness on abstention-matching responses.

Primary 5-judge panel only. Mean score on the 192 abstention-matching responses:

Judge	Mean on abstentions
Sonnet 4.6	1.14
GPT-5.4	1.17
Haiku 4.5	1.29
GPT-4o	1.34
Opus 4.6	1.41

Spread: 0.27 points between strictest (Sonnet) and most lenient (Opus). No judge reaches the rubric-1 floor on average. The 5-judge primary average (1.27) smooths this cross-judge variance without eliminating it.

D.3.4 Length-score correlation.

Pearson correlation between response length (character count) and 5-judge primary score, across the 1,599 low-baseline responses:

Slice	n	r	Interpretation
All responses	1,599	0.26	Modest positive, driven almost entirely by C5.
C5 (baseline, no context)	312	0.604	Strong positive. Longer baseline responses score higher.
C2a (Spec only)	351	0.14	Near zero.
C4 (facts alone)	312	0.01	Zero.
C4a (facts plus Spec)	312	−0.01	Zero.
C2c (wrong Spec)	312	0.500	Strong positive. Wrong-Spec responses resemble C5 on the length-score axis.

The effect is strongest in C5 (r = 0.604) and recurs, attenuated, in C2c (r = 0.500). Both are conditions without a ground-truth representation of the subject: C5 has no context at all, and C2c has a randomly-drawn other subject's specification. In both, longer responses (containing hedging, adjacent-fact recitation, disambiguation offers) score higher than short refusals. Conditions that do carry a correct specification (C2a, C4, C4a) show near-zero length correlation. The direction of the

bias pushes the measured C5 and C2c means upward, which shrinks the measured Spec-vs-no-Spec gap relative to the true gap. That the length signal persists in C2c, but not in C2a or C4a, is the cleanest evidence that length inflation is a property of the baseline-scoring regime rather than of any specific condition: when judges cannot verify against a correct representation, they partial-credit verbose output.

D.3.5 Ultra-high-score validity.

Ultra-high responses are those scoring 4.5 or above on the 5-judge primary. Length comparison:

Response class	Mean length (chars)	Notes
Ultra-high (score 4.5 or above)	2,790	Not length-inflated.
Mid-range (2.5 to 3.5)	2,829	Baseline comparison.
Low (score below 2.0)	2,087	n = 795. Shorter than both ultra-high and mid-range; confirms length inflation is a low-end partial-credit phenomenon, not a high-end one.

Ultra-high responses are not longer than mid-range responses. Length inflation is a low-end phenomenon, not a universal one. The hypothesis that "ultra-high responses equal length-inflated responses" is rejected by this comparison.

D.3.6 Implications for reported effects.

Both rubric-handling effects (abstention over-credit at the low end, length inflation in C5) pull the measured C5 baseline upward. This shrinks the measured Spec-vs-baseline gap. The true effect size for the population of relevance is likely somewhat larger than the +0.89 mean lift reported in §4.1. The paper reports the measured number rather than a length-corrected one to keep the pre-locked analysis plan intact. A differentiated rubric that scores abstention as its own dimension, and a length-controlled scoring protocol, are both flagged as follow-up in §7.

D.4 Per-judge score matrices

Per-subject by per-judge score means for C5 (baseline) and C4a (facts plus Spec) conditions are derivable from the raw per-judge JSON files under `results/global_<subject>/*_judgments_<judge>.json` (and `results/hamerton/` for Hamerton). The slice-level picture is already reported in §3.3.3 (calibration) and §4.6.2 (5-judge vs 7-judge sensitivity), which together establish that directional agreement is tight (Spearman ρ 0.86 to 0.93) while absolute magnitude varies (Krippendorff α 0.659 5-judge, 0.535 7-judge).

Full per-subject by per-judge by per-condition mean-score matrix for the 14 main-study subjects across the 5 gradient conditions (C5, C2a, C2c, C4, C4a):

Each cell is the per-judge mean score across all behavioral-prediction questions for a (subject, condition, judge) triple. Judges abbreviated: H=Haiku 4.5, S=Sonnet 4.6, O=Opus 4.6, 4o=GPT-4o, 5.4=GPT-5.4, gF=Gemini 2.5 Flash, gP=Gemini 2.5 Pro. 5m = 5-judge primary mean, 7m = 7-judge mean.

"n/a" indicates missing judge-condition coverage (most commonly: Gemini judges not run on C2c or C4 for some subjects; see §4.6.2 on 5-judge vs 7-judge coverage).

Subject	Cond	H	S	O	4o	5.4	gF	gP	5m	7m
Hamerton	C5	1.23	1.15	1.36	1.33	1.21	1.28	1.16	1.26	1.25
	C2a	2.72	2.13	3.05	2.67	2.59	3.49	3.50	2.63	2.88
	C2c	1.38	1.36	1.69	1.38	1.44	2.03	2.56	1.45	1.69
	C4	2.26	2.26	2.87	2.33	2.41	2.64	3.11	2.43	2.55
	C4a	2.69	2.38	3.26	2.87	2.64	3.87	3.92	2.77	3.09
Sunity Devee	C5	1.03	1.00	1.05	1.05	1.00	1.08	n/a	1.03	1.03
	C2a	2.41	1.79	2.56	2.15	2.41	3.49	n/a	2.27	2.47
	C2c	1.28	1.13	1.38	1.38	1.28	1.72	n/a	1.29	1.36
	C4	2.59	2.15	2.74	2.44	2.38	3.54	n/a	2.46	2.64
	C4a	2.46	2.13	2.59	2.49	2.38	3.58	n/a	2.41	2.61
Ebers	C5	1.00	1.00	1.05	1.05	1.00	1.13	n/a	1.02	1.04
	C2a	1.49	1.31	1.82	1.56	1.51	3.08	n/a	1.54	1.79
	C2c	1.41	1.10	1.38	1.44	1.26	2.44	n/a	1.32	1.50
	C4	2.21	1.59	2.26	2.03	2.03	3.15	n/a	2.02	2.21
	C4a	2.26	1.62	2.31	2.10	2.08	3.67	n/a	2.07	2.34
Fukuzawa	C5	1.64	1.44	2.00	1.64	1.64	2.46	n/a	1.67	1.80
	C2a	2.18	1.97	2.79	2.41	2.41	3.56	n/a	2.35	2.56
	C2c	1.85	1.49	2.38	2.21	1.74	2.97	n/a	1.93	2.11
	C4	2.95	2.28	3.00	2.54	2.59	3.95	n/a	2.67	2.88
	C4a	2.85	2.26	3.21	2.77	2.82	4.03	n/a	2.78	2.99
Seacole	C5	1.69	1.49	1.92	2.00	1.77	2.24	n/a	1.77	1.85
	C2a	2.44	2.08	2.72	2.56	2.62	3.44	n/a	2.48	2.64
	C2c	1.33	1.26	1.69	1.49	1.38	1.87	n/a	1.43	1.50
	C4	3.13	1.95	2.82	2.69	2.54	3.51	n/a	2.63	2.77
	C4a	2.74	2.13	2.82	2.72	2.56	3.72	n/a	2.59	2.78
Bernal Díaz	C5	1.72	1.31	1.87	1.85	1.74	2.64	1.67	1.70	1.83
	C2a	2.18	1.85	2.62	2.38	2.31	3.62	2.75	2.27	2.53
	C2c	1.64	1.54	1.90	2.00	1.87	2.82	3.14	1.79	2.13
	C4	2.59	1.87	2.67	2.46	2.46	3.51	3.40	2.41	2.71
	C4a	2.28	2.18	2.79	2.56	2.59	3.46	3.60	2.48	2.78
Keckley	C5	2.00	1.56	1.85	1.82	1.97	2.28	n/a	1.84	1.91
	C2a	2.38	1.90	2.69	2.51	2.64	3.72	n/a	2.43	2.64
	C2c	1.28	1.21	1.54	1.46	1.28	2.23	n/a	1.35	1.50

Subject	Cond	H	S	O	4o	5.4	gF	gP	5m	7m
	C4	2.64	1.95	2.46	2.49	2.41	3.46	n/a	2.39	2.57
	C4a	2.33	2.03	2.56	2.56	2.69	3.54	n/a	2.44	2.62
Yung Wing	C5	2.08	1.62	1.97	1.90	1.82	2.36	n/a	1.88	1.96
	C2a	2.28	1.95	2.51	2.26	2.08	3.31	n/a	2.22	2.40
	C2c	2.15	2.00	2.33	2.21	2.31	2.97	n/a	2.20	2.33
	C4	2.15	1.82	2.36	2.18	2.13	2.90	n/a	2.13	2.26
	C4a	2.38	2.13	2.74	2.38	2.36	3.18	n/a	2.40	2.53
Bābur	C5	1.79	1.41	1.79	2.10	1.69	2.90	2.53	1.76	2.03
	C2a	1.92	1.49	2.23	2.21	1.69	2.87	3.53	1.91	2.28
	C2c	1.23	1.03	1.23	1.23	1.13	1.64	1.14	1.17	1.23
	C4	2.18	1.59	2.10	2.26	2.03	3.36	3.06	2.03	2.37
	C4a	2.13	1.77	2.18	2.15	1.82	3.18	3.47	2.01	2.39
Cellini	C5	2.64	1.90	2.54	2.51	2.31	3.46	n/a	2.38	2.56
	C2a	2.31	2.26	2.85	2.62	2.69	3.62	n/a	2.54	2.72
	C2c	1.79	1.59	1.90	2.00	1.79	2.59	n/a	1.82	1.94
	C4	2.44	2.03	2.74	2.51	2.38	3.56	n/a	2.42	2.61
	C4a	2.56	2.28	2.69	2.56	2.54	4.13	n/a	2.53	2.79
Zitkala-Ša	C5	2.62	1.85	2.46	2.38	2.38	3.90	n/a	2.34	2.60
	C2a	2.15	1.64	2.21	2.05	2.10	3.00	n/a	2.03	2.19
	C2c	1.82	1.36	1.87	1.69	1.56	2.23	n/a	1.66	1.76
	C4	2.41	1.72	2.31	2.08	2.00	3.28	n/a	2.10	2.30
	C4a	2.00	1.74	2.26	2.10	2.00	3.49	n/a	2.02	2.26
Rousseau	C5	2.59	1.85	2.62	2.64	2.49	3.72	n/a	2.44	2.65
	C2a	2.77	2.23	3.00	2.95	3.10	4.05	n/a	2.81	3.02
	C2c	1.74	1.59	2.44	1.90	1.90	3.28	n/a	1.91	2.14
	C4	2.44	1.90	2.59	2.36	2.33	3.46	n/a	2.32	2.51
	C4a	2.72	2.03	2.64	2.49	2.79	3.74	n/a	2.53	2.74
Augustine	C5	3.00	1.95	2.64	2.69	2.64	3.79	2.90	2.58	2.80
	C2a	2.62	1.85	2.72	2.69	2.51	4.08	4.36	2.48	2.97
	C2c	2.10	1.64	2.21	2.41	2.21	3.90	3.33	2.11	2.54
	C4	2.77	2.08	2.62	2.85	2.49	4.18	4.67	2.56	3.09
	C4a	2.72	2.10	2.79	2.97	2.90	4.56	4.50	2.70	3.22
Equiano	C5	2.92	2.28	2.95	2.97	2.72	3.74	n/a	2.77	2.93

Subject	Cond	H	S	O	4o	5.4	gF	gP	5m	7m
	C2a	2.44	1.97	2.77	2.56	2.54	3.90	n/a	2.46	2.70
	C2c	1.92	1.51	2.36	2.26	1.82	3.18	n/a	1.97	2.18
	C4	2.62	2.23	2.67	2.49	2.15	3.64	n/a	2.43	2.63
	C4a	2.51	2.00	2.67	2.67	2.26	3.82	n/a	2.42	2.65

Total cells: 14 subjects x 5 conditions x 9 columns = 630. Source: raw per-judge JSON files under `results/global_<subject>/*_judgments_<judge>.json` (global subjects) and `results/hamerton/` (Hamerton), aggregated via `scripts/_emit_full_judge_matrix.py`.

The matrix is 14 subjects x 5 conditions x (7 per-judge columns + 5-judge primary mean + 7-judge mean) = 70 rows x 9 columns = 630 cells. Subject rows follow the C5-baseline ordering used in §4.1 (lowest baseline first). Empty Subject cells continue the previous subject's rows. Gemini Pro "n/a" entries reflect that Gemini Pro was run as a sensitivity judge only on a subset of subjects (§4.6.2); those cells were never populated. Franklin is not included in this matrix because the Franklin control condition set does not align to the C5 / C2a / C2c / C4 / C4a labels used in the global-subject run; Franklin's judgments are reported in §4.2 and are stored under `results/franklin_legacy_20260411/analysis/`. Source: `scripts/_emit_full_judge_matrix.py`, which aggregates from `results/global_<subject>/*_judgments_<jud` and `results/hamerton/`.

D.5 Example verbatim responses at each rubric anchor

Verbatim-response examples at rubric anchors 1-5 for one representative subject (Hamerton) are in §3.3 as part of the rubric definition. Examples at anchor crossings are developed in §4.1 (Examples A, B, C on Ebers, Bernal Díaz, Seacole). Three illustrative paired (C5, C4a) per-question excerpts for each of the 14 main-study subjects are collected in Appendix E. The raw response JSON files (`results/global_<subject>/results_v2.json`) contain every response verbatim alongside its 5-judge primary score.

Appendix E. Per-subject worked examples

Appendix E. Per-subject worked examples. The 14 per-subject worked examples (one per main-study subject) are in `docs/supplementary/appendix_E_per_subject_worked_examples.md` in the public repository. Each subject has three illustrative paired (C5, C4a) per-question response excerpts under the 5-judge primary score, selected by the largest C4a minus C5 panel-mean Δ within each subject. The excerpts are organized to illustrate the cross-anchor interpretation rule (§3.3.1) at the subject level and to make concrete the multi-anchor crossings discussed in §4.1.1. Per-subject baselines (C5) range from 1.02 (Ebers) to 2.77 (Equiano); per-question Δ_C4a values in the supplementary file range from +0.40 to +4.00. Full response artifacts are at `results/<subject>/results_v2.json` and `results/hamerton/results.json`.

Appendix F. Benchmark Scope Analysis

This appendix develops, benchmark by benchmark, the scope differences summarized in §2.1 between prior work on memory and personalization benchmarks and what this paper measures. The load-bearing point in each case is the same: representational accuracy, operationalized as behavioral prediction on held-out reasoning situations, is not what these benchmarks evaluate. None of them is wrong on its own axis. None of them is a substitute for the test in this paper.

F.1 LongMemEval

Reference. Wu et al., ICLR 2025 (Wu et al., 2025).

Task. Evaluate long-term memory in chat assistants across multiple sessions. Five capability dimensions: single-session memory, multi-session reasoning, temporal reasoning, knowledge updates, and abstention.

Scoring. Question-answering accuracy, with held-out facts embedded across session history and queried in a later session. Answers are compared against ground-truth factual targets drawn from the same session history the system ingested.

Training / test protocol. Conversation history is ingested; the system is then queried with fact-recall questions whose answers are present in the ingested history. The test is whether the memory system can surface the correct facts at retrieval time.

What it measures. Fact recall across long context windows. A secondary axis tests whether the system correctly abstains when the answer is not in the conversation history.

What it does not measure. Whether the memory system's representation of a specific person captures how that person reasons. Every LongMemEval target is a fact that was literally said in the conversation; no target is a held-out behavioral pattern.

Published range. Memory systems reported in the 68% to 85% range depending on provider, model, and benchmark variant (cited in §1.1 and §2.2). Specific numbers per system are in the papers and vendor reports.

Relationship to this paper's test. Orthogonal axis. Our battery targets held-out behavioral patterns that were never literally said in the training half of the corpus; every question is backward-designed to answer only from patterns, not from retrievable content. A system that ranks at the top of LongMemEval can still sit near the rubric floor on our battery, and a system that ranks low on LongMemEval (for example, Base Layer's retrieval substrate) can contribute on our battery through the specification rather than through retrieval.

F.2 PersonaGym

Reference. Samuel et al., Findings of EMNLP 2025 (Samuel et al., 2025).

Task. Evaluate persona fidelity in conversational agents. Given a described persona, measure whether the model maintains that persona across a conversation.

Scoring. Persona-consistency metrics over multi-turn conversation. LLM-judge evaluation of whether the model's voice, stated preferences, and surface-level claims remain consistent with the described persona.

Training / test protocol. A persona is described (occupation, background, preferences, mannerisms). The model is prompted to roleplay the persona across a dialogue. Evaluation is whether the dialogue responses remain internally consistent with the persona description.

What it measures. Persona presentation fidelity. Can the model stay in character on the described dimensions.

What it does not measure. Whether the model accurately predicts how the person described by the persona would respond to new situations. A persona-fidelity system can maintain voice without

ever accurately anticipating decisions. A representationally accurate system can shift voice (for example, from formal prose to casual conversational register) while continuing to predict accurately on behavioral questions.

Published best-number. Top PersonaScore of 4.51 ± 0.08 on a 1-5 scale (GPT-4.5), out of 10 evaluated LLMs spanning 200 personas and 10,000 questions; bottom of the range was 3.64 ± 0.57 (Claude 3 Haiku). Notably, GPT-4.1 and LLaMA-3-8b tied on PersonaScore despite a large capability gap (Samuel et al., 2025). Scoring is on persona-consistency metrics, not held-out behavioral prediction; these numbers are not directly comparable to this paper's rubric means on the 1-5 behavioral-prediction scale.

Relationship to this paper's test. Both measure something that is sometimes called "personalization," but the axes are different. PersonaGym is surface-presentation consistency; our battery is transfer of the subject's interpretive patterns to unseen situations. Our rubric does not credit voice-matching alone (score 2: "generic, not subject-specific"); it requires capturing the behavioral pattern the subject exhibited in the held-out passage.

F.3 AlpsBench

Reference. Xiao et al., 2026 (Xiao et al., 2026).

Task. Evaluate whether explicit memory mechanisms improve preference-aligned and emotionally resonant responses in conversational settings.

Scoring. Preference-alignment scoring and emotional-resonance scoring on conversational responses, both LLM-judged against reference targets derived from user preference data.

Training / test protocol. A conversational agent is seeded with a user's preference history (via an explicit memory mechanism, or as a baseline without one). The agent responds to new prompts. Responses are scored on preference-alignment and emotional-resonance metrics.

What it measures. Whether explicit memory makes conversational responses more aligned with stated preferences and more emotionally appropriate.

What it does not measure. Whether memory mechanisms enable the model to predict the user's behavior in unseen reasoning situations. Preference alignment and behavioral prediction are related but distinct: a system can match preferences on immediate response choices without having a representation of the user's reasoning that transfers to situations the system has never seen.

Central finding. AlpsBench's central empirical result is that explicit memory mechanisms improve recall but do not guarantee more preference-aligned or emotionally resonant responses. This is independently arrived at and complementary to our own finding. They find the gap in preference alignment; we find it in behavioral prediction. Both point in the same direction: recall-solving is insufficient for what memory is ultimately for.

Relationship to this paper's test. Adjacent. Same motivating intuition (recall improvement does not transfer to downstream behavioral properties), different downstream property measured.

F.4 Twin-2K

Reference. Toubia et al., 2025 (Toubia et al., 2025).

Task. Behavioral prediction at scale. 2,058 participants each answered a large-scale survey, and the system predicts each participant's responses on held-out survey items given a persona constructed from their other survey answers.

Scoring. Distance metric on Likert-scale items (numeric distance between predicted response and actual response, aggregated per participant and per item).

Training / test protocol. For each participant, a subset of survey answers is used to author a persona. The persona is served to a model as context. The model predicts the participant's answer on the held-out survey items. Distance between predicted and actual response is the score.

What it measures. Behavioral prediction on survey-response interpolation. Does a machine-readable transcript of one half of a participant's survey predict the other half.

What it does not measure. Behavioral prediction on open-ended reasoning situations. Twin-2K's held-out items are additional Likert responses from the same survey form; the test is interpolation across a structured response distribution. Our held-out items are open-ended behavioral predictions on unseen autobiographical passages; scoring is via rubric on response content, not distance on a numeric scale.

Relationship to this paper's test. Closest prior work on the behavioral-prediction axis. Three structural differences remain load-bearing:

1. **Task format.** Twin-2K: Likert interpolation. This paper: open-ended behavioral prediction with free-text answers.
2. **Persona construction.** Twin-2K: machine-readable transcript of the participant's own prior survey responses, served as-is. This paper: an authored Behavioral Specification composed of three interpretive layers plus a composed brief. The Twin-2K persona is raw data; our specification is compressed interpretation.
3. **Held-out distance.** Twin-2K: the held-out items are drawn from the same structured survey instrument. This paper: the held-out items are drawn from autobiographical text the representation has never seen, in a different form than the training half (different chapters, different situations).

Twin-2K measures whether a model can interpolate a person's survey distribution from other survey responses. Our battery measures whether a representation of how a person reasons transfers to new situations the representation has never seen. Both are legitimate tests of behavioral prediction; neither is a substitute for the other.

An earlier exploratory Base Layer run against Twin-2K's battery produced positive results on that task format, but we do not report those numbers as a formal benchmark comparison because the experiment used a prior iteration of our pipeline, and the task targets are substantively different (see §2.1).

Published best-number. Top individual-level accuracy of 71.72% on held-out survey items using a text-persona representation served to GPT-4.1-mini (Toubia et al., 2025). Human test-retest reliability on the same instrument was 81.72%, putting the top twin at 87.67% of the human ceiling. Random-guess baseline was 59.17%. Aggregate-level replication: the Twin-2K twins reproduced results from 6 of 10 behavioral-economics experiments, with systematic divergences on medical decision-making and political attitudes. The 71.72% accuracy is on Likert interpolation, which is a structurally different task from our rubric-scored free-text behavioral prediction.

F.5 LoCoMo

Reference. Maharana et al., ACL 2024 (Maharana et al., 2024).

Task. Conversational memory quality over long multi-session dialogues.

Scoring. Fact-recall questions over ingested dialogue history. Similar scope to LongMemEval but focused on conversational-memory substrates specifically.

Training / test protocol. A long multi-session conversation is ingested; the memory system is queried on specific facts from earlier sessions.

What it measures. Long-dialogue recall accuracy.

What it does not measure. Behavioral reasoning. LoCoMo targets are literal recalls from session history.

Published range. LoCoMo paper baselines (Maharana et al., 2024): GPT-4-turbo 32.1% overall, GPT-3.5-turbo 22.4%, GPT-3.5-turbo-16K 37.8%, best RAG configuration 41.4%; human performance 87.9%. Memory-system claims on LoCoMo, detailed in §2.2: Mem0g variant 68.44 with GPT-4o-mini (Chhikara et al., 2025); Mem0 production algorithm 91.6 self-reported with open-sourced evaluation harness; Letta 74.0 with GPT-4o-mini; earlier Zep claim of 84 publicly disputed by Mem0 (see §2.2 dispute note). The methodology disagreement between vendors remains unresolved; §2.2 treats these single-number comparisons with explicit caution.

Relationship to this paper's test. The benchmark the four memory systems (Zep, Letta, Mem0, Supermemory) compete on. §2.2 uses these results as context for the memory-system landscape. Our paper is orthogonal: we do not evaluate memory systems on LoCoMo; we evaluate their behavioral-prediction performance on held-out autobiographical passages with and without the Behavioral Specification added.

F.6 MemOS and related systems-level benchmarks

Reference. Systems-level memory benchmarking literature, including MemOS and adjacent evaluations. See §2.2 for the memory-systems landscape.

Task. Evaluate memory-layer infrastructure choices (storage substrate, retrieval algorithm, consistency properties) rather than memory-quality outcomes.

Scoring. Varies. Typically: retrieval latency, throughput, consistency guarantees, scalability benchmarks.

What it measures. Infrastructure properties.

What it does not measure. Representational accuracy, persona fidelity, or preference alignment. Systems-level benchmarks do not evaluate the quality of the representation the memory layer produces; they evaluate the mechanics of how that representation is stored and served.

Relationship to this paper's test. Different layer of the stack. Our paper evaluates what gets stored and why; systems-level benchmarks evaluate how well it is stored and served. Both matter for deployed personal-AI systems. The specification and the memory-layer infrastructure compose: our §4.4 Mem0 / Letta / Zep / Supermemory / Base Layer results show the specification adding on top of each infrastructure choice, not replacing it.

F.7 What no prior benchmark measures

Pulling the per-benchmark analysis together, the axis that representational accuracy sits on is not covered by any prior benchmark:

1. **Test data the system has not seen.** LongMemEval, PersonaGym, and LoCoMo target content the system has ingested. Twin-2K's held-out items are drawn from the same structured instrument. Our battery's held-out passages are from unseen chapters in a different narrative register than the training half.
2. **Open-ended behavioral prediction rather than structured-format scoring.** Twin-2K is the closest comparison; it is Likert-format rather than open-ended.
3. **Representation of how a person reasons, not what they said or prefer.** PersonaGym tests voice consistency; AlpsBench tests preference alignment; LongMemEval / LoCoMo tests fact recall. None tests transfer of interpretive patterns.

This is the gap the paper's battery targets. The battery is not a replacement for any of the above. It is a test of a different property: whether a representation of a specific person enables a model that has never seen the person's held-out reasoning to anticipate it accurately.

F.8 Persona-input depth comparison across benchmarks

The §2.1 footnote `[^twin2k-persona-size]` calls out that Twin-2K's persona input is much deeper than PersonaGym's one-line descriptor. This subsection collects persona-input depth across the benchmarks named in §2.1 and Appendix F so the comparison is concrete. "Persona-input depth" is the total token volume of the participant- or subject-specific representation served to the model at inference time, measured on the input that an evaluated system actually consumes.

Benchmark	Persona-input form	Approximate input depth (tokens)	Notes
LongMemEval (Wu et al., ICLR 2025)	Multi-session conversation transcript ingested before query	varies, ~thousands per session × 5 sessions	The "persona" is the user's accumulated conversational history. Depth depends on configured session count; the standard `_s` configuration runs ~5 ingested sessions before query.
LoCoMo (Maharana et al., ACL 2024)	Long multi-session dialogue ingested before query	varies, ~10K-100K depending on session count	Similar to LongMemEval in form: ingested dialogue is the "persona." Single-session and multi-session test variants.

Benchmark	Persona-input form	Approximate input depth (tokens)	Notes
PersonaGym (Samuel et al., Findings of EMNLP 2025)	One-line descriptor (e.g., *"You are a 45-year-old skeptical accountant from Toronto"*)	≈ 20–50	Surface attributes only; no behavioral depth. The benchmark is designed to test consistency with a thin descriptor across multi-turn conversation.
AlpsBench (Xiao et al., 2026)	User profile of preferences and prior emotional-context exchanges	varies, ~hundreds to low thousands	Profile is a concatenation of stated preferences and selected prior dialogue snippets. Designed to test preference-aligned response generation rather than reasoning transfer.
Twin-2K (Toubia et al., 2025)	Full survey-response transcript or condensed summary	`persona_text` ≈ 32,000; `persona_summary` ≈ 3,750	Two persona-input variants are reported in the released code. `persona_text` is the participant's full prior-question-and-answer transcript; `persona_summary` is a model-condensed form.
Base Layer (this paper)	Behavioral Specification (anchors / core / predictions composed into a unified brief)	≈ 7,000	Compressed-interpretation form. Per-subject specification sizes range from ≈ 4,000 (Hamerton, smallest corpus) to ≈ 9,000 (longest-corpus subjects). Compression ratios vs. source corpus span 5× to 80× (§4.2).

Reading. The benchmarks span roughly three orders of magnitude in persona-input depth, from

PersonaGym's ~50 tokens to Twin-2K `persona_text`'s ~32,000. Two structural points fall out of the comparison.

First, persona-input depth alone does not predict what a benchmark measures. Twin-2K's full transcript is far deeper than this paper's specification, but the task is Likert interpolation across the same survey instrument; what is being measured is whether a model can complete a participant's response distribution, not whether it captures their interpretive patterns on situations the representation has never seen.

Second, the form of the persona input matters more than the size. PersonaGym's one-line descriptor and Twin-2K's full transcript are at opposite ends of the depth scale, and both are "raw" inputs in different senses (a descriptor of surface attributes versus a serialized record of prior responses). The Behavioral Specification at ~7,000 tokens sits in the middle of the depth range but is structurally distinct: it is an authored compression of how the subject reasons, not a transcript or a descriptor. The compression-curve evidence in §4.2 shows that this structural distinction is what makes the form behave the way it does in cross-format behavioral prediction.

Sources: footnote-cited papers and the released code or schema where applicable. PersonaGym descriptor length is from the canonical example in Samuel et al. §3.1; AlpsBench profile composition is from Xiao et al. §3.2; Twin-2K depth values are from `persona_text` and `persona_summary` columns in the released dataset; Base Layer specification depth is per-subject from `data/global_subjects/<subject>/spec/brief.md`.

Appendix G. Letta Stateful-Agent: Exploratory Case Study (full)

Body summary in §4.5. This appendix retains the full method, per-subject results, robustness checks, content analysis, and caveats from the original §4.5 in v9 / earlier drafts of v10.

This section is a post-hoc exploration, not a replication or a headline finding. N=3 subjects (Hamerton, Ebers, Bābur), one Letta version, one response model (Claude Haiku), a 40-question battery per subject. The intent is to characterize what Letta's stateful-agent architecture produces when invoked directly, and how that compares at matched response model to Base Layer's unified-brief variant. It is not an attempt to establish that two systems reach a common target.

> **Headline result on the small sample tested (5-judge primary):** Letta's self-edited memory block scores higher than Base Layer's unified-brief variant on all 3 subjects at matched response model. Hamerton **3.10 vs. 2.96** (Δ +0.14), Ebers **2.76 vs. 1.72** (Δ +1.05), Bābur **2.42 vs. 1.88** (Δ +0.54). A robustness rerun against Base Layer's full layered stack preserves direction (Δ +0.27 / +1.21 / +0.38). The matched-model gap is largest at the mid-corpus subject (Ebers) and smaller at both endpoints; with three data points the shape is consistent with a corpus-size band where the self-edited block is most effective, with degradation as the block grows, or with insufficient interpretive content when the corpus is small. Multi-subject replication is flagged as the highest-priority external falsification (§7.5).

Letta is the one commercial memory system in the study whose architecture supports an alternative to retrieval at query time. Alongside the archival retrieval path tested in §4.4, Letta agents maintain a persistent memory block that the agent itself rewrites during ingestion. This is the stateful-agent design from the original MemGPT paper. It is architecturally distinct from retrieval-based memory:

the representation is authored by the agent over the course of reading the source corpus, rather than chunked and indexed for later retrieval. §4.5 examines what that produces on a small set of subjects, with the caveats above. Multi-subject replication across the full gradient, multiple response models, and a comparison against the Base Layer full layered stack (rather than the compressed variant used here) are flagged as follow-ups in §7.5.

Test design. A fresh Letta agent was initialized and fed the training half of each subject's corpus turn-by-turn. The agent was allowed to self-edit its memory block during ingestion, its native MemGPT behavior. After ingestion, the resulting memory block was extracted and served as context to Claude Haiku 4.5, the response model used throughout the main study. The behavioral-prediction battery was the main-study battery. Three subjects were tested, spanning a 9× corpus-size range:

Subject	Source corpus	Corpus size (words)	Letta block size (chars)
Hamerton	Philip Gilbert Hamerton, *An Autobiography* (training half)	25,231	22,472
Ebers	Georg Ebers, *The Story of My Life* (training half)	48,161	68,413
Bābur	Bābur, *Bābur-nama* (training half)	222,742	335,349

The direct comparison: Letta's stateful-path memory block fed to Haiku, vs. Base Layer's full-stack specification fed to the same Haiku, on the same battery and judge panel. Both are interpretive representations delivered as context; the test isolates the representation itself.

Methodological note on the Base Layer condition served here. The Base Layer side of this matched-rerun loaded the unified brief variant (a ~7K-character synthesized document served as a single artifact) rather than the full layered stack (anchors + core + predictions + brief) that §4.4's controlled and native C2a / C3 conditions use. The unified brief is more compressed on referential detail than the layered stack. A layered-stack rerun on these three subjects would likely narrow the Letta-over-BL gap; whether it narrows to parity or reverses is not measured. The table column header below reflects this: the Base Layer side is the unified brief variant.

Result (5-judge primary: Haiku, Sonnet, Opus, GPT-4o, GPT-5.4).

Subject	Letta block → Haiku	BL unified brief → Haiku	Δ (Letta − BL)
Hamerton	3.10	2.96	**+0.14**
Ebers	2.76	1.72	**+1.05**
Bābur	2.42	1.88	**+0.54**

On all three subjects tested, Letta's stateful-path block, served to the same response model as the

Base Layer unified brief, produces a higher per-subject mean score than the unified brief. Both representations land well above the retrieval-only baseline at matched response model (§4.4 Letta archival Δ_spec for these subjects: Hamerton near parity with Base Layer retrieval, Ebers +0.31, Bābur near-null).

Judge-panel robustness. The 7-judge sensitivity aggregate (Hamerton +0.093, Ebers +0.746, Bābur +0.232; see `docs/research/letta_stateful_matched_rerun.md` Part 7 appendix) preserves direction on all three subjects. The 5-judge primary values are larger than the 7-judge values on Ebers and Bābur by +0.30 and +0.31 points respectively, because the two Gemini judges were inflating Base Layer scores relative to the calibrated core on those subjects. Excluding Gemini from the aggregate (the paper's 5-judge primary convention; §3.3.3 and §4.6.2) therefore widens the Letta-over-BL gap rather than narrowing it. Hamerton is the exception (5-judge Δ +0.14 vs. 7-judge +0.09), where Gemini inclusion slightly narrowed the gap rather than widening it. In all three cases, the Letta-block-outperforms-BL-Spec direction is stable across panels.

Compression behavior: divergence at large corpora.

Letta's memory block grew roughly linearly with source corpus size. At the largest subject (Bābur), Letta's API began rejecting ingestion requests at approximately 333,000 characters. After 22 consecutive failed ingestion attempts, the final block measured 335,349 characters. Letta's declared block-size metadata limit is 100,000 characters, unenforced in practice; the effective ceiling on the server side appeared to be a different API-level limit around 333K.

At the ceiling, the block contained **25.4% verbatim sentence duplication** on Bābur, compared to 0% duplication on Hamerton and 0% on Ebers. The self-editing agent rewrites content it has already written when pressed against the ingestion limit, rather than compressing or summarizing. The representation carries corpus-derived narrative at scale but does not preserve the compression property that makes large corpora tractable.

Semantic-similarity duplication. A sentence-embedding analysis (post-hoc, this paper; `scripts/analyze_letta_semantic_duplication.py`, MiniLM-L6-v2, sentence-pair cosine $\geq$ threshold) shows that the verbatim figure understates the duplication. The self-editing agent paraphrases prior sentences as well as repeating them. On Bābur, 73.8% of sentences have a near-duplicate at cosine $\geq$ 0.80, 56.1% at $\geq$ 0.85, 41.4% at $\geq$ 0.90, and 35.2% at the strict $\geq$ 0.95 threshold (paraphrase-level matches). Ebers shows minor near-paraphrasing (11.5% / 3.3% / 1.1% / 0.5% across the same thresholds). Hamerton shows none at any threshold above 0.80. The pattern matches the verbatim direction. Sample matches at $\geq$ 0.95 on Bābur include `"Emotional Resilience in Governance: Bābur's personal reflections..."` paired with `"Emotional Resilience in Leadership: Bābur's reflections on challenges..."` (cosine 0.957): same template, slight rewording. The agent's abstention behavior tracks this duplication gradient: on Bābur (most degraded), Letta abstains on 17.9% of held-out questions vs 0% for the Spec; on Ebers, 10.3% vs 5.1%; on Hamerton (least degraded), the rate inverts to 7.7% vs 10.3% — consistent with adaptive recognition of block degradation rather than ceiling-induced confusion. The duplication within the block does *not* propagate as surface-syntactic leakage into responses: a per-question 5-gram overlap test against held-out passages returns 0.0% on every single question for both Letta and Spec (§4.5 mechanism paragraph). The duplication is a within-block artifact of self-editing at the ceiling; the response-level mechanism for Letta's lift is named-entity grounding plus content-confidence. Full per-threshold duplication data at `docs/research/letta_semantic_duplication_20260501.json`; per-subject abstention

decomposition at `docs/reviews/letta_vs_spec_abstention_20260507.md`; per-question leakage analysis at `docs/research/letta_vs_spec_leakage_analysis_20260507.md`.

Base Layer's compose step keeps the full-stack specification at 34,000-40,000 characters across the same corpus-size range. At Hamerton, the two representations are the same order of magnitude in size; at Bābur, the Base Layer specification is roughly one-tenth the size of the Letta block. The two systems are prediction-band compatible at small corpora; they diverge on compression at large ones.

What the ceiling means for deployment. Served on every query, a 335,000-character Letta block costs roughly 84,000 tokens of context. At current frontier pricing this is materially more per-query cost than the Base Layer specification's ~7,000 tokens (~37,000 characters), and it exceeds the context window on the smaller-context models still common in production (128K token windows struggle when the block alone is two-thirds of the budget, before any conversational state). The duplication observed at the ceiling combines 25.4% verbatim sentence repetition with substantial semantic near-paraphrasing (35.2% of sentences at cosine ≥ 0.95, 56.1% at ≥ 0.85). The block would be functionally much smaller with a deduplication pass. Conservatively (one-of-each-pair removal at ≥ 0.85), roughly 30% of the block is removable; aggressive cluster-collapse deduplication at the same threshold could reach a 50% reduction, taking the block from ~335K to ~170K characters at preserved content. A semantic-similarity deduplication pass on the self-edited block is a tractable post-processing step that this study does not run but recommends. For production deployment, the ceiling, the verbatim duplication, and the additional semantic duplication together argue for representation compactness as a first-class design constraint, not a nice-to-have.

What this exploration does and does not show.

On N=3 subjects, with one response model and one Letta version, Letta's stateful-path block and Base Layer's unified-brief variant both land above retrieval-only context at matched response model, in a similar prediction band. This is consistent with (though does not establish) the idea that the behavioral-specification target is reachable by representation-production mechanisms outside offline-authored retrieval composition. Establishing that would require multi-subject replication across the full gradient, multiple response models, and a comparison against Base Layer's full layered stack rather than the unified-brief variant tested here. All three are flagged in §7.5.

What the exploration does show is the shape of the engineering tradeoff between the two paths. They differ in how the representation is produced (offline authoring vs. online self-editing), in what it carries (interpretive scaffolding vs. corpus-derived narrative at higher referential density; see content comparison below), and in how it scales (bounded compression vs. an ingestion ceiling observed at ~333K characters on the largest corpus we tested). These are tradeoffs to characterize, not a resolved comparison.

Content comparison: what each representation retains.

To test whether Letta's higher matched-model score comes from preserving original corpus text the response model could cite, we ran a post-hoc content analysis on the three subjects. The strong form of that hypothesis is refuted. Neither representation is a quote library. Checking what fraction of consecutive five-word sequences in each representation also appears verbatim in the training corpus (a standard overlap check), both representations score under 1%: the Letta block ranges 0.0-1.0% depending on subject, the Base Layer specification scores 0.0% on all three. The same check for

consecutive ten-word sequences gives under 0.1% for both. Both representations are LLM-generated rewrites of the corpus in the writing model's own voice, not verbatim extracts.

A refined version of the hypothesis does hold, with the magnitude smaller than first reported. The two representations differ in **referential density**: Letta's rolling summary retains more unique proper nouns, dated events, and named secondary characters than Base Layer's §4.5 specification, and the gap scales with corpus size. On Bābur (the largest corpus), Letta's block carries 416 unique capitalized named-entity tokens vs. Base Layer's 65, a ratio of about **6×**. On Ebers (mid-size), the counts are 53 vs. 34, a ratio of about **1.5×**, closer to parity. Base Layer, by construction, compresses episodes into cross-cutting behavioral patterns with fewer surface referents; the pipeline explicitly anonymizes the subject during authoring and compresses corpus-level specifics into dimensional axioms. Letta's stateful-agent path preserves more of the referential surface while also encoding behavioral patterns. The referential-density gap is real but corpus-dependent rather than uniformly an order of magnitude.

Both representations produce responses that outperform retrieval-only context at matched response model, but they diverge on referential detail. On battery items that reward specific-event recall, Letta has more named entities to cite. On items that reward principled interpretation across episodes, Base Layer's dimensional axioms compete directly. The §4.5 matched-model gap may be attributable in part to the referential-density difference rather than to the self-editing process itself. A Base Layer variant that retains named entities inside the same dimensional scaffold would separate the two effects. Flagged in §7.

Replication as the load-bearing next step. The three-subject comparison reported here is not a claim that alternative representation-production architectures reach the interpretive-representation target. It is a case study with direction but not power. Multi-subject replication across the full 14-subject gradient (layered-stack Base Layer vs. Letta stateful, both anonymized to match, multiple response models) is the highest-priority external falsification we can run on §4.5, and is flagged as such in §7.5. If that replication closes the gap at parity, §4.5's direction holds on a wider sample. If it reverses, §4.5's direction was corpus-specific.

Full content analysis at `docs/research/` (see `_content_analysis_results.json` and the N=3 per-subject breakdown). The methodological note on the Base Layer condition is now hoisted above the result Table at the top of this section.

Caveats.

- N = 3 subjects on this path. Extending across the full 14-subject gradient would let the comparison speak to the population-of-relevance level, not only a selected set of corpus sizes. Flagged in §7.5.
- One response model (Haiku) on both conditions. The comparison is tested at matched response model; whether it holds at other response models is an open question.
- Letta's 333K-character ingestion ceiling is a hard architectural constraint in the current release. For small corpora the two representations are interchangeable in prediction behavior; for large corpora the ceiling is material.
- Base Layer condition used the unified `spec.md` variant for the main §4.5 table. A robustness rerun with the full layered stack (anchors + core + predictions + brief, name-restored to match the §4.5 naming convention) preserves direction on all three subjects (Δ_Letta−BL = +0.27 / +1.21 / +0.38 on Hamerton / Ebers / Bābur; full report at

`docs/research/_letta_rerun/fullstack_named/RESULTS.md`). The gap widens at the two smaller corpora and narrows at Bābur, consistent with a Pattern 2 (over-theorization) effect on small corpora rather than a content-volume effect at large corpora. Direction is invariant across both Base Layer Spec forms.

- **Naming asymmetry.** Letta's stateful-agent path ingested the named source corpus and wrote a memory block that references the subject by name throughout. Base Layer's authoring pipeline strips the subject's name during specification authoring (§3.7 anonymization step); the §4.5 comparison restores the name at the surface level only (string substitution on the composed artifact). The two sides of the comparison therefore differ in whether the subject's name is load-bearing during representation production vs. only at serving time. Flagged as a methodological gap in §7.5.

Raw data and scripts. Letta stateful matched-rerun data at `docs/research/_letta_rerun/{subject}_judgmen` Generation and scoring scripts live in the same directory as a numbered chain (`20_run_c2a_named.py`, `40_judge_responses.py`, `60_rerun_gpt54_letta.py`, `70_compute_5judge_primary.py`); see the `README.md` inside `docs/research/_letta_rerun/`. Full characterization of block content, duplication behavior, and API responses in `docs/research/letta_stateful_deep_read.md` and `docs/research/letta_stateful_matched_rerun.md`.

Appendix H. Glossary

Defined terms used as terms of art throughout the paper.

5-judge primary panel. The locked judge aggregation for headline numbers. Aggregation rule: per-judge per-question score → per-judge per-subject mean → panel mean across {Haiku 4.5, Sonnet 4.6, Opus 4.6, GPT-4o, GPT-5.4}. See §3.3.3.

7-judge sensitivity panel. The 5-judge primary plus Gemini 2.5 Flash and Gemini 2.5 Pro, reported as a sensitivity check. See §3.3.3.

Anchors / Core / Predictions. The three layered artifacts comprising a Behavioral Specification. Anchors: short axiomatic claims about how the subject reasons. Core: connects anchors into coherent reasoning patterns. Predictions: derives forward-looking decisions from the core. A composed brief sits above all three. See §3.7.

Behavioral prediction. The operational test for representational accuracy. Given a situation drawn from text the model has never seen, the model generates how the subject would respond; the response is scored against the subject's own verbatim response on a 1-5 interpretive rubric. See §1.1, §3.3.

Behavioral Specification. A static document of approximately 7,000 tokens that extracts and encodes a person's behavioral patterns. Composed of three layered artifacts (anchors, core, predictions) plus a unified brief. Layered above memory-system retrieval as an interpretive structure. See §1.1, §3.7.

Cross-anchor interpretation rule. A fractional delta between two conditions that crosses an integer rubric anchor (1, 2, 3, 4, or 5) reflects a categorical shift in the underlying response

distribution. A delta that stays inside a single integer band is a within-category shift and a weaker claim. See §3.3.1.

Interpretation. In this paper, the human-side property: the way a specific person processes facts and experiences into judgments, decisions, and reactions. The property the Behavioral Specification is designed to mirror. See §1.1.

Multi-anchor crossing. A single question whose 5-judge primary mean shifts across two or more integer rubric bands when the condition changes. Crossings can span two bands (e.g., $1 \rightarrow 3$) or, more rarely, three bands (e.g., $1 \rightarrow 4$ or $2 \rightarrow 5$). The strongest categorical signal the rubric detects. See §3.3.1, §4.2.1.

Personalization (this paper's sense). Representing the interpretive layer that sits beneath stated preferences and biographical facts: how a specific person organizes experience, what they treat as evidence, what reasoning patterns they apply across new situations. Distinguished from surface-level responsiveness to stated preferences (dietary restrictions, communication style) or stored facts about the user (location, occupation, history), which are downstream artifacts of the interpretive layer rather than the layer itself. See §2 lede.

Refusal (abstention). A response in which the model declines to predict because the available context does not support a prediction. Distinct from a substantively wrong prediction; under the current rubric both score at the lowest anchor (1). Detection in the validity-audit script uses phrase patterns ("no specific information," "I cannot confirm," "would need additional context"). See §3.3.6.

Representational accuracy. The AI-side property: how faithfully a model's internal representation of a specific person captures that person's interpretive patterns. Operationalized via behavioral prediction on held-out reasoning situations. See §1.1, §3.1.

Specification-effect claim. When a Behavioral Specification is served to the model as context, the model's responses shift in the direction of the subject's demonstrated behavioral patterns, and that shift registers as a measured increase in representational accuracy against held-out passages from the same subject. The claim is directional, not a claim of new model capability or absolute correctness. See §3.3.4.

Tier 1 / Tier 2. Tier 1 is the main study: Haiku 4.5 response model across all 14 subjects, every condition, Haiku-generated batteries. Tier 2 is the cross-provider directional probe: Sonnet 4.6 and Gemini 2.5 Pro response models on 3 subjects (Ebers, Yung Wing, Zitkala-Ša) with GPT-5.4-regenerated batteries. See §3.6, §4.6.1.

Wrong-Spec control. A deliberately mismatched Behavioral Specification served in place of the correct one. Two variants: **v1** (adversarial fixed derangement maximizing cultural and temporal distance; aggregate Δ -0.25) and **v2** (seed-fixed random derangement; aggregate Δ $+0.15$). See §1.3, §3.2, §4.3.