CHRONODOSSIER: LONG-FORM PERSONAS AS TEMPORAL MEMORY REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon agents must recover a person's current state from stored records, but recovery grows less reliable as those records lengthen and their facts change. An agent can retrieve the right passage and still answer from an obsolete or incomplete state. Many long-form memory benchmarks evaluate records only after construction, which leaves unclear whether such an error arose when the record was written or read. We treat a long-form persona dossier as a memory representation of a known latent state and evaluate it both against that state and through the models that read it. We present CHRONODOSSIER, a benchmark of 311 synthetic dossiers of 1k–10k words written by seven generators, each from a dated timeline, with every passage linked to its life phase. Intrinsic metrics measure coverage, cross-phase leakage, and collapse across personas; extrinsic metrics measure how well five models recover the encoded state from the full dossier or from retrieved passages across dossier lengths and as facts change, and attribute explicitly supported failures to retrieval, packaging, or answer formation. How a memory is written shapes how well it can be read. Scoping each passage writer to its own life phase removes all detected cross-phase leakage without a measured loss of coverage, and replay items authored on staged dossiers pass 11 percentage points less often on one-pass dossiers of the same latent states. Reading then fails in ways that retrieval metrics do not register: top-three retrieval reaches the correct life phase for every unique frontier item, and only 3 of 128 frontier failures on explicit facts begin with a retrieval miss, yet GPT-5.1 and Claude Sonnet 4.6 score 11 to 34 points lower on 10k-word than on 1k-word dossiers, depending on how evidence is packaged, against 4 points with the full dossier. To the best of our knowledge, CHRONODOSSIER is the first benchmark to trace construction defects and reading failures in a long-form persona memory to individual passages and life phases.

1 INTRODUCTION

An agent serving one person across many sessions answers, plans, and advises from what it has stored about that person. In agentic use, a memory error changes what the agent does: answering from a superseded role, an outdated routine, or an unsupported premise acts on the wrong state. Reliability requires preserving intended distinctions when a record is written and recovering relevant evidence when a later decision needs it. Time complicates both stages: a fact can hold in one life phase and not the next. Most existing benchmarks measure recovery. LongMemEval tests extraction, multi-session reasoning, temporal reasoning, updates, and abstention (Wu et al., 2025); AMemGym evaluates interactive state evolution (Cheng et al., 2026); and long-context suites show that nominal context capacity does not ensure task performance (Yen et al., 2025). In these settings, the record usually arrives fully formed. Synthetic-persona generation exposes the preceding problem: diverse seeds can yield realizations that converge, leak facts between sections, or blur which phase governs a claim.

We study a long-form persona dossier as a memory representation. Each dossier encodes a structured latent state, invariant identity plus a dated timeline of facts, routines, and decision rules, and that state is known because the dossier is generated from it. *Intrinsic* metrics compare the representation with its source state and with other personas' representations; *extrinsic* metrics ask whether a model reading the representation recovers the state across dossier lengths, as access narrows to retrieved passages, and as the governing life phase changes. We hypothesize that longer representations are harder to use, most of all under retrieval. The study asks three questions:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

- Which intended distinctions does the constructed representation encode, and does it keep personas and life phases apart?
- How well do models recover the encoded state across dossier lengths, as access narrows, and as facts change over time?
- Where do remaining failures arise, and do changes to the encoding, the evidence, or the model's weights reduce them?

We present CHRONODOSSIER, a benchmark with three parts. It *constructs* the representation, turning a structured temporal seed into a provenance-preserving artifact graph whose views encode the same state. A *metric suite* (Section 4) evaluates that representation intrinsically and extrinsically; gold-section labels and answer keys attribute explicitly supported failures to retrieval, packaging, or answer formation. *Replay* tests how well five models carry the representation into answers from the full dossier (FULL) or retrieved passages (RETRIEVE), under alternative encodings of the same evidence, and after parameter-efficient adaptation, which tests whether a model can learn to use it.

The results are consistent with the length hypothesis for top-three retrieved access. At 10k words, frontier RETRIEVE accuracy in the main configuration is 0.705 to 0.745 of FULL accuracy, the drop shrinks when the same passages arrive as longer excerpts, and time-dependent explicitly supported items trail timeless ones under retrieval although retrieval selects the correct phase. The losses concentrate in particular question types: artifact-and-workflow questions are hardest, compound questions lose most under retrieval, and most retrieval loss falls on explicit facts re-asked after scripted history. Construction choices change which distinctions are realized and how reliably models recover them. For an open-source reader, re-encoding retrieved sections as versioned state records does not raise accuracy, parameter-efficient adaptation lowers inference accuracy while explicitly supported accuracy stays flat, and deeper retrieval raises recall without a robust gain. The claim is specific to this task and artifact format; human realism, psychological validity, and deployment utility remain outside it.

2 RELATED WORK AND POSITIONING

Persona resources and behavioral evaluation. PERSONA-CHAT established profile-conditioned dialogue (Zhang et al., 2018), and Dialogue NLI and later personalization work made persona consistency and implicit inference measurable (Welleck et al., 2019; Madotto et al., 2019; Zhou et al., 2023). Persona Hub scales persona data; RoleLLM/RoleBench and OpenCharacter support role-play (Ge et al., 2024; Wang et al., 2024a; 2025). Among behavioral evaluations, InCharacter uses psychological interviews (Wang et al., 2024b); PersonaGym and CharacterBench use persona-relevant environments and targeted questions (Samuel et al., 2025; Zhou et al., 2025); TimeChara tests knowledge at a selected narrative point (Ahn et al., 2024); and PersonaArena, SPASM, and PersonaForge study context-rich or long-horizon behavior (Shi et al., 2026; Luo & Laban, 2026; Tong & Zou, 2026). CHRONODOSSIER examines the representation upstream of these behaviors: a long-form dossier with inspectable provenance and claims that change across explicit life phases. Replay accuracy measures artifact-internal recoverability and does not measure human simulation fidelity; Appendix Table 7 compares representation scale across these resources.

Long-context, memory, and temporal evaluation. Long-context benchmarks test retrieval, reasoning, and robustness as evidence grows. HELMET shows that task coverage and evaluation design materially affect model rankings (Yen et al., 2025). LongMemEval decomposes long-term conversational memory into information extraction, multi-session reasoning, temporal reasoning, updates, and abstention (Wu et al., 2025); LoCoMo evaluates question answering and event summarization over long conversations (Maharana et al., 2024); and AMemGym adds structured state evolution and on-policy interaction (Cheng et al., 2026). MSC extends persona dialogue across sessions (Xu et al., 2022), and PersonaMem and PrefEval test whether models track evolving user profiles and preferences (Jiang et al., 2025; Zhao et al., 2025). Agent memory systems decide what to store and retrieve (Park et al., 2023; Packer et al., 2023), and the granularity of memory or retrieval units shapes what is found (Pan et al., 2025; Chen et al., 2024). TimeChara and temporal RAG benchmarks isolate point-in-time fidelity and retrieval over indexed evidence (Ahn et al., 2024; Chen et al., 2025; Cao et al., 2026). Their inputs evolve as dialogue, news, or event streams. CHRONODOSSIER fixes the record in advance: replay selects among co-existing life phases within one dossier, which isolates representation realization from the accumulation and retrieval of later history.

Synthetic-data audits and documentation. Synthetic-data quality is often summarized by similarity, diversity, or model-based judgments. Sentence embeddings (Reimers & Gurevych, 2019) can expose semantic concentration that lexical overlap misses, and recent work evaluates synthetic data through downstream effects alongside surface quality (Fan et al., 2026). Data statements, datasheets, and model cards motivate release procedures that keep realized outputs inspectable (Bender & Friedman, 2018; Gebru et al., 2021; Mitchell et al., 2019). CHRONODOSSIER applies both views to one representation: intrinsic audits and downstream replay share the same records.

3 PERSONA DOSSIERS AS MEMORY REPRESENTATIONS

3.1 LATENT STATE AND ITS ENCODINGS

A persona's latent state s has invariant identity fields and a timeline of phase-bounded facts, routines, constraints, and decision rules. Seeds vary along 11 audited axes: age range, gender, ethnicity/background, region, urban form, climate context, education, wealth, job, specialization, and work schedule. This operational target does not claim demographic completeness.

Construction encodes s as a representation exported in five linked views (Figure 1): (1) the long dossier used by FULL; (2) a compact card of facts, values, heuristics, boundaries, and timelines; (3) section records with tags and time scope; (4) phase capsules sliced from multi-phase sections; and (5) repair logs. Every passage keeps its persona, section, and phase identifiers; a replay miss therefore links back to the frozen item, retrieved sections, card, capsule, and any pre-export repairs. Here, "graph" refers to this joinable provenance structure; no graph representation is learned. Retrieved evidence can be re-encoded in two further forms. *Versioned state records* list each changing fact's current and superseded values under one key; a *resolved-state packet* selects the same records but states the value that holds at the queried year and keeps only the immediately superseded version.

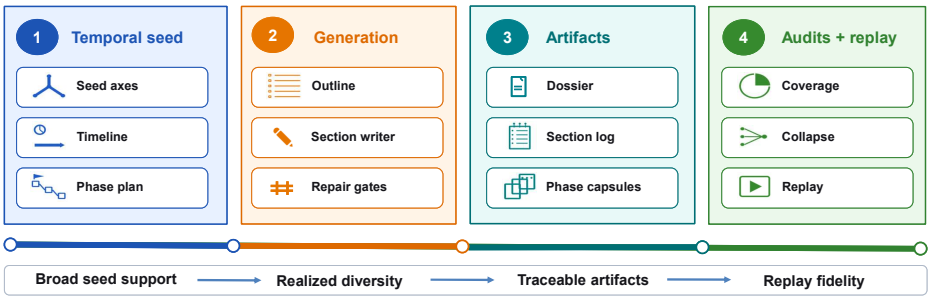


Figure 1: CHRONODOSSIER joins construction and downstream use. A temporal seed is sliced into phase-local payloads; accepted sections become dossiers, cards, and phase capsules; audits and replay share their identifiers. FULL supplies the full dossier and RETRIEVE three section passages.

Worked record. One released 10k-word realization shows how the views encode one latent state and how a replay miss traces back through them. The seed places Daichi Lee in a mid-size riverfront suburb with a riverfront trail network and four distinct seasons. His card states that "Daichi Lee is a 43-year-old man as of 2025" who "works as a senior construction manager in a remote-first construction-tech startup," holds a master's degree in construction management completed part-time between 2017 and 2019, and "uses standardized decision documentation templates containing context, status, decision needed, options, recommendation, and next actions." Its decision rules default to conservative schedule and cost estimates when information is thin, keep a recurring checklist of scope, safety, cost, schedule, and learning value, and run pre-mortems before critical-path milestones. One explicitly supported item asks for the suburb, the trail network, and the four seasons. The FULL answer passes. One RETRIEVE turn retrieves sections `s11`, `s1`, and `s3` but omits the trail and climate details; another retrieves `s11`, `s6`, and `s1`, includes all required facts, and passes. The chain separates topical retrieval from supplying every required fact.

3.2 Construction procedure and controls

Before prose generation, an outline planner assigns sections to phase IDs, year bounds, and required topic tags. The section writer receives invariant identity and only the phase-local seed fields relevant to the section, plus a compact continuity state from accepted sections. Structural and temporal gates test tags, chronology, phase scope, and length; novelty gates compare a draft with accepted text. Failed sections are repaired or regenerated. Two matched controls isolate construction choices. The *payload control* gives matched section requests either the full seed or only the section's own tag-scoped fields. The *flat control* replaces timeline-first stage-local generation with one-pass generation while holding seeds, timelines, nominal lengths, frozen replay items, target model (GPT-5.1), and judge fixed.

Data provenance and human involvement. All records are synthetic. Seven generators write them from analyst-defined seeds: GPT-5.1, GPT-5 mini, Claude Sonnet 4.6, Claude Haiku 4.5, Gemma 4 26B-A4B (Gemma Team, 2026), Llama 3.1 70B Instruct (Grattafiori et al., 2024), and Qwen3 32B (Qwen Team, 2025) (Appendix Table 12). Short names omit versions; GPT denotes GPT-5.1. The resource includes no study participants or collected real-person profiles. Outputs are validated, repaired, and exported without manual rewriting. Humans define seed axes, rubrics, and the protocol; they also provide the limited adjudication in Section 6 and the phase-rubric calibration in Appendix C.1. Requested, completed, frozen, and rated counts remain separate.

4 Evaluating a Memory Representation

Table 1: Metric suite. Intrinsic metrics need no reader; extrinsic metrics require one.

Metric	Type	Definition
Quality	Intrinsic	Consensus judge score on an eight-criterion document rubric
Collapse	Intrinsic	Mean nearest-neighbor cosine between dossier embeddings (Top1), its 95th percentile, and pairs above 0.8
Cross-phase leakage	Intrinsic	Sections with an exact multiword seed value from outside their scope
Coverage	Intrinsic	Share of permitted in-scope seed anchors that the section realizes
Recoverability	Extrinsic	Pass rate of a reader under an access condition, over all rated turns
Source-item transfer	Extrinsic	Recoverability on items authored from a different realization of the same state
Length retention	Extrinsic	RETRIEVE accuracy divided by FULL accuracy within a length bucket
Temporal gap	Extrinsic	Phase-linked minus timeless accuracy on explicitly supported items
Failure attribution	Extrinsic	Explicitly supported failures by cause: retrieval, packaging, answer formation

4.1 Intrinsic metrics

Table 1 lists the metric suite. GPT-5.1 and Claude Sonnet 4.6 judges share one document rubric (Appendix Table 9) covering internal and temporal consistency, seed fidelity, specificity, readability, bias avoidance, retrieval readiness, and length control. Gemini 3 Pro Preview is added as a third judge when their scores differ by over 15 points, unless both are below 70. Model judges are measurement instruments; their outputs are not gold standards (Liu et al., 2023; Zheng et al., 2023). Collapse uses all-MiniLM-L6-v2 dossier embeddings: high Top1 means that distinct latent states produced similar representations, and the 95th percentile exposes clusters that the mean hides; exact lexical 5-gram overlap adds a surface view. Leakage counts exact multiword seed phrases (anchors) copied from an axis outside the requested scope. Phase packets, one per capsule, also receive an ordinal A–F temporal-distinctness tier, calibrated against human raters in Appendix C.1.

4.2 Readers and extrinsic metrics

Replay uses 32 held-out GPT-5.1 source personas, eight per 1k, 2.5k, 5k, and 10k target-length bucket, and five readers: GPT-5.1 and Sonnet (the *frontier* targets) and the open-source Qwen, Llama, and Gemma. GPT-5.1 authors item definitions once from the seed, card, timeline, dossier, and section excerpts under a JSON schema; accepted items are frozen before target-model runs. The 431 unique definitions form 520 rated turns per condition. Retention ladders re-ask anchor items and time-jump episodes revisit earlier-phase items, reusing some definitions. Of 8,880 turns across 96 episodes,

8,360 are scripted history and 520 are rated live turns (Appendix Table 13); replay therefore measures answers after controlled histories and excludes fully model-generated conversation.

Items fall into three families. *Surface-faithful* (SF) answers are explicitly supported; *inferred-faithful* (IF) answers require inference from stored values, heuristics, or patterns; and *not-in-persona* (NIP) requests require refusal because the premise is unsupported, fabricated, or privacy-invasive. Each item stores its phase, key or rubric, and gold section labels. SF items use answer keys with fallbacks to atomic NLI and to the judge; NIP items use refusal checks; IF items receive a Boolean judgment from the GPT-5.1 judge. A pass must satisfy the family rule and any phase constraint. Replay scores the eligible user-visible answer; an unterminated reasoning block or empty final is a nonanswer and fails. Judge payloads omit reader identity and access mode, although evidence can reveal the mode.

FULL supplies the entire dossier. RETRIEVE supplies an *evidence packet* of the top three section passages, each capped at 550 characters, from hybrid retrieval over all-MiniLM-L6-v2 embeddings and lexical, heading, tag, and year matches. The open-source RETRIEVE runs in the main comparison ran without the embedder, which was not loaded, and used the other matches alone; all other reported RETRIEVE runs use it (Appendix D). A *query planner* builds the query from the current request and the two previous user turns: frontier targets write a structured query, and open-source targets use a fixed heuristic planner, adopted after none of 77 planning calls in a Qwen3 32B pilot returned a parseable query within 180 tokens. The item, reader, scoring rule, and history stay fixed across access modes. Neither condition isolates retrieval or serves as an oracle: retrieval-augmented generation couples evidence selection with answer formation (Lewis et al., 2020), and long-context models use relevant evidence unevenly across positions and distractors (Liu et al., 2024; Yen et al., 2025). Retrieval diagnostics record gold phase hit@3 and section hit@3 (a top-three passage comes from the item's phase or a labeled gold section) and phase purity. Failure attribution applies each frozen SF answer key to the labeled source sections and to the realized packet: a failure is assigned to retrieval when no labeled section is selected, to packaging when a labeled section is selected but the packet fails the key, and to answer formation when the packet passes the key.

4.3 INTERVENTIONS ON EVIDENCE, ENCODING, AND WEIGHTS

The *gold-section condition* supplies only the sections labeled for each item; 408 of the 431 definitions carry section labels, all resolving to stored sections. The *fixed-total* policy distributes at most 1,650 characters across the selected passages, separating passage allocation from context growth. The depth and adaptation contrasts each score both arms in one run on the 32-persona panel and cache judge decisions for identical responses; the format contrast compares runs that share items and retrieved sections. The *depth contrast* raises k from 3 to 12 at the 1,650-character total for GPT-5.1, Sonnet, and Gemma with the heuristic planner. The *format contrast* gives Gemma the same retrieved sections as raw text or as versioned state records; on the eight-persona screen, it also compares versioned records with the resolved-state packet at matched budgets. The *adaptation contrast* compares untuned Gemma with a rank-16 QLoRA adapter (Hu et al., 2022; Dettmers et al., 2023) trained by supervised fine-tuning (SFT), under byte-identical prompts with versioned state records. The adapter was trained for one epoch on 62 personas disjoint from every evaluation panel, with passing GPT-5.1 full-dossier responses as targets and answer keys or refusals as fallbacks.

Every rated turn contributes to the headline rates, including nonanswers. Availability is reported separately from answer-conditional quality. Proportions use Wilson intervals; overall rates and paired contrasts of frozen turns use length-stratified, persona-clustered bootstrap intervals (Appendix E).

5 RESULTS

5.1 INTRINSIC: WHAT THE REPRESENTATION ENCODES

Collapse and quality diverge. Document quality and collapse capture different properties of the 311 completed dossiers (Table 2). The highest-quality generator (Sonnet) is not the least collapsed (Haiku), and Qwen's low judged quality coincides with a cluster of similar dossiers: Qwen is the only generator with pairs above the 0.8 alert threshold. Its worst alert links two distinct dossiers whose prose begins with near-identical instructions to "rewrite the Biography section"; repeated task framing drives the similarity, and the personas do not share the implicated fact. Qwen stays the most collapsed in both views under equal-size subsampling. Surface deduplication alone would

miss such concentration: matched Wikipedia biographies are less concentrated in embedding space than the synthetic documents at comparable lexical 5-gram overlap (Appendix C). All generators export the same artifact families under timeline-first planning, stage-local writing, and live repair; the documents remain model-specific pipeline outputs and are not matched semantic realizations.

Table 2: Intrinsic metrics by generator. C/R is completed/requested. Quality is the mean consensus score: the judge-pair mean, or the three-judge median when Gemini is added. Top1 is mean nearest-neighbor cosine, p95 its upper tail, and Alerts counts pairs above 0.8.

Generator	C/R	Quality↑	Top1↓	p95↓	Alerts↓
GPT-5.1	94/100	84.69	0.5797	0.6723	0
GPT-5 mini	37/40	82.70	0.5794	0.6402	0
Sonnet 4.6	35/40	**88.06**	0.5365	0.6064	0
Haiku 4.5	37/40	81.92	**0.5232**	**0.6029**	0
Gemma 4 26B-A4B	32/40	63.41	0.6274	0.6889	0
Llama 3.1 70B	37/40	53.46	0.5696	0.6593	0
Qwen3 32B	39/40	47.77	0.7773	0.8364	12

Leakage and transfer. Scoping each writer's payload removes detected cross-phase leakage without a measured loss of coverage. With the full seed, 29 of 100 sections copy an anchor from outside their scope; with tag-scoped payloads, none do, and in-scope coverage changes by $+0.016$ (95% CI $[-0.019, +0.051]$). Construction also changes transfer: frozen items pass on 95.5% of 202 turns against stage-local dossiers and 84.2% against one-pass dossiers of the same states (-11.3 points, persona-bootstrap 95% CI $[-23.1, -3.4]$). Because items were authored against staged dossiers, this measures transfer for the whole construction package, not stage-locality alone.

5.2 Extrinsic: recovering the encoded state

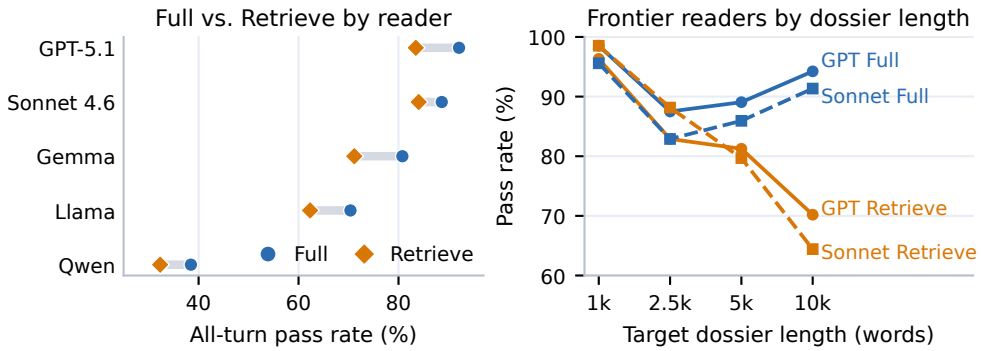


Figure 2: Left: all-turn replay pass rates after user-visible answer extraction, with nonanswers counted as failures. Right: frontier readers by dossier length. Retrieved-context accuracy is lower for longer dossiers although the top three passages include the correct phase for every unique frontier item.

Because Figure 2's left panel and Table 3 count nonanswers as failures, their access gaps also reflect interface completion; the right panel shows only frontier readers, which answer every turn.

Length retention. Consistent with the hypothesis, frontier RETRIEVE accuracy declines across the length buckets, which hold different personas: it is 26 points lower for GPT and 34 for Sonnet on 10k-word than on 1k-word dossiers, against 4 under FULL; at 10k words, RETRIEVE keeps 0.745 and 0.705 of FULL accuracy (Table 4). The size of the decline depends on evidence packaging. With the heuristic planner and identical top-three sections, the decline is 24.6 and 29.2 points with 550-character excerpts and 12.7 and 10.7 with longer excerpts under a 1,650-character total (cross-run comparisons, Appendix Table 17). The loss arises after retrieval reaches the right part of the dossier: the top three passages contain the item's phase for every unique frontier item and a labeled section for 94.7% (GPT) and 97.3% (Sonnet) of them (Appendix F). The diagnostics cannot tell whether the excerpt omitted a needed fact or the reader failed to use it.

Table 3: Recoverability (% pass) over 32 held-out personas, 520 rated turns per row, by score family. All-turn rates count nonanswers as failures. Open-source RETRIEVE rows retrieve without embeddings (Section 4.2). Appendix Tables 14 and 15 give counts and CIs.

Access	Target	Overall	SF	IF	NIP
FULL	GPT	**92.12**	94.17	87.13	88.14
FULL	Sonnet	88.65	91.94	89.11	67.80
RETRIEVE	GPT	83.46	80.83	91.09	86.44
RETRIEVE	Sonnet	**84.04**	83.61	89.11	77.97
FULL	Qwen	38.46	46.94	19.80	18.64
RETRIEVE	Qwen	32.31	34.44	33.66	16.95
FULL	Llama	70.38	87.22	39.60	20.34
RETRIEVE	Llama	62.31	73.06	37.62	38.98
FULL	Gemma	80.77	82.22	77.23	77.97
RETRIEVE	Gemma	71.15	68.06	79.21	76.27

Table 4: Extrinsic robustness of frontier readers. Retention divides RETRIEVE by FULL pass rates per length bucket (Appendix Table 17). The temporal gap is phase-linked minus timeless SF accuracy over 115 and 245 items, a post-hoc split (Appendix Table 18).

	Length retention (RETRIEVE/FULL)				Temporal gap (pts)	
Reader	1k	2.5k	5k	10k	FULL	RETRIEVE
GPT	0.978	0.947	0.912	0.745	+2.2	−12.7
Sonnet	1.031	1.063	0.927	0.705	−3.5	−11.7

Temporal gap. Time-dependent facts lose more under retrieval: under RETRIEVE, phase-linked SF items trail timeless ones by 12.7 points for GPT and 11.7 for Sonnet, against small, mixed-sign FULL gaps (Table 4). Because retrieval selects the correct phase, the post-hoc split points to a temporal evidence-access problem; small IF and NIP cells preclude a uniform claim across families.

Open-source readers. For open-source readers, returning an answer is itself a failure source. Qwen exposes an eligible answer on only 273 of 520 FULL and 274 of 520 RETRIEVE turns, often stopping inside an unterminated reasoning block, and nonanswers account for most of its failures; on the SF turns it answers under FULL, it passes 86.2%, close to Llama's 87.2% (Appendix Table 16). Llama and Gemma answer every turn, and answer extraction changes their rates by less than one point. Gemma leads open-source overall and answer-conditional performance; Llama leads open-source SF but passes under 40% of IF and NIP turns. Gemma is therefore the only open-source reader with consistent performance across item families, and the interventions in Section 5.4 use it. Answer formation, interface completion, and memory remain separate failure sources.

5.3 WHICH QUESTIONS RESIST RECOVERY

Table 5 breaks the 520 core turns down by question type; three types stand out.

Artifacts and workflows are hardest. Among explicitly supported items, questions about the artifacts and workflows a persona relies on (for example, which planning or tracking templates a persona used as of a given year) have the lowest pass rates for every reader under both access modes. For the frontier readers, they are the only tabulated SF type below 90% under FULL.

Compound questions lose most under retrieval. Two-fact questions lose 8.0 to 16.0 points under RETRIEVE, against 3.6 to 7.3 for atomic ones, consistent with fragment-integration failure.

Retrieval loss sits in re-asked facts. Most of the frontier SF loss under retrieval falls on facts that retention episodes re-ask after scripted history: these episodes hold 47 of GPT's 55 and 37 of Sonnet's 43 SF turns that pass only under FULL, while time-jump episodes change by two turns per frontier reader. Even identity and context items, the largest SF type, fall 9 to 13 points under RETRIEVE for all three readers from near-ceiling FULL accuracy. False-premise handling varies by reader: even under FULL, Sonnet passes 58.8% of false timeline or role claims and 69.7% of prompt injections.

Table 5: Pass rates (%) by question type on the 520 core turns. Subcategories with fewer than 15 turns are omitted. GPT and Sonnet use the runs of Table 3; Gemma uses the current-pipeline same-run control, separate from Table 3. Bold: paired exact McNemar $p < 0.05$ for FULL against RETRIEVE.

		GPT		Sonnet		Gemma	
Question type	n	FULL	RETRIEVE	FULL	RETRIEVE	FULL	RETRIEVE
SF: identity and context	266	97.4	**84.6**	97.7	**86.5**	98.9	**89.8**
SF: phase-specific history	46	97.8	**82.6**	91.3	89.1	93.5	82.6
SF: artifacts and workflows	34	79.4	**58.8**	67.6	55.9	70.6	**47.1**
IF: decision heuristics	65	86.2	87.7	86.2	92.3	61.5	47.7
IF: values and goal trade-offs	15	93.3	100.0	93.3	93.3	60.0	66.7
NIP: false timeline or role claim	17	82.4	64.7	58.8	70.6	88.2	88.2
NIP: prompt injection	33	97.0	97.0	69.7	78.8	100.0	100.0
Atomic questions	395	93.9	**86.6**	90.4	86.8	89.4	**84.1**
Compound questions	125	86.4	**73.6**	83.2	75.2	78.4	**62.4**
Retention episodes (all SF)	256	93.4	**77.3**	92.2	**80.9**	91.4	**82.8**
Time-jump episodes	136	89.7	88.2	88.2	86.8	71.3	67.6
Adversarial episodes	128	92.2	90.6	82.0	87.5	93.8	**82.8**

Three patterns recurred in a qualitative review of failed turns. A *cross-phase application* uses an obsolete rule despite identifying the right topic. In *false-premise compliance*, the answer accepts an incorrect date, role, or affiliation that the evidence contradicts. A *fragment integration failure* follows one retrieved passage and overlooks a constraint found in another. Topical retrieval can therefore succeed in all three cases while the final answer still fails.

5.4 CHANGING THE ENCODING, THE EVIDENCE, OR THE WEIGHTS

On the 32-persona panel, no intervention changes overall accuracy significantly, and no packet encoding reaches full-dossier accuracy. Only within-row changes are interpreted.

Encoding. Re-encoding Gemma's retrieved sections as versioned state records leaves accuracy unchanged (-1.1 points, $p = 0.58$) and lowers NIP accuracy by 11.9 ($p = 0.016$); every packet encoding stays 7.9 to 9.2 points below the full dossier (Appendix Table 19). Stating the value that holds at the queried year helps on the eight-persona screen: a resolved-state packet passes 122 of 154 turns against 111 for versioned records at 8,000 characters ($p = 0.019$).

Depth. Increasing k from 3 to 12 changes the evidence mix even though the total packet length is fixed: gold-section recall rises from 0.64 to 0.97, and the amount of gold-section text per turn falls by half. Gemma's IF accuracy rises by 14.9 points ($p = 0.028$, CI $[+3.8, +26.8]$), a gain that falls to 5.9 points ($p = 0.24$) without text after its end-of-turn marker (Appendix H). Its SF and overall accuracy do not change. GPT and Sonnet already pass 92 and 97 of the 101 IF turns, respectively, and neither changes significantly (Table 6). Their gains are confined to cases where top-3 retrieval found no gold section (GPT $+9/-0$, $p = 0.004$).

Weights. The SFT adapter lowers IF accuracy by 21.8 points ($p = 0.0001$, CI $[-33.7, -9.4]$) and raises NIP accuracy by 11.9 ($p = 0.039$), with SF and overall accuracy unchanged. Lost IF answers are mostly incomplete: adapted IF answers are half as long, and the judge cites omissions for 22 of the 27. On the eight-persona screen, none of ten adapters from five recipes exceeds the untuned IF count or changes overall accuracy significantly.

6 ANALYSIS AND DIAGNOSTIC CONTROLS

Judge substitution. Most comparisons survive a change of judge. Rescoring every GPT-5.1-judged turn with Sonnet 4.6 reproduces 90.4% of decisions and nine of ten model/access comparisons, while absolute rates rise by 2.9 to 4.8 points, especially for rubric-only IF items. Blinded human review of 24 stratified judge disagreements found each sampled gold label and rubric valid. This sample of disputed items gives neither a corpus-wide accuracy estimate nor a human upper bound.

Table 6: Paired pass counts on the 520 rated turns of Table 3; headers give turn counts. Depth rows fix total evidence at 1,650 characters and use the heuristic planner; adaptation uses versioned state records in both arms. Each row is one run with cached judge calls. Bold: exact McNemar $p < 0.05$.

Contrast	IF (101)	SF (360)	NIP (59)	Overall (520)
GPT-5.1, k 3→12	92→96	315→312	54→48	461→456
Sonnet 4.6, k 3→12	97→94	314→302	43→46	454→442
Gemma, k 3→12	**48→63**	298→295	57→59	403→417
Gemma, untuned→adapted	**45→23**	309→316	**50→57**	404→396

Planner substitution and regrading. Aggregate frontier rates are stable across planners: a fixed heuristic planner changes frontier RETRIEVE accuracy by -2.5 (GPT) and -4.6 points (Sonnet), and both intervals include zero (Appendix F). Judge decisions are not fully deterministic: regrading 154 byte-identical responses changed 6 decisions, and the Section 5.4 contrasts cache judge calls.

Failure attribution. Frontier SF failures rarely begin with a retrieval miss: only 3 of GPT's 69 and none of Sonnet's 59 are assigned to retrieval. Most arise later, in packaging, when a labeled section is selected but its excerpt drops the keyed answer, or in answer formation, when the packet holds the answer and the reader still fails (Appendix Table 20). Readers differ where evidence is incomplete. On byte-identical $k = 3$ packets, GPT, Sonnet, and Gemma pass 97 to 98% of turns whose packet satisfies the key; on the remaining turns, GPT and Sonnet pass 71% and Gemma 61%. These counts attribute failed turns under the deterministic key contract and do not causally partition every answer.

7 LIMITATIONS AND SCOPE

Judge reliability. Model judges score document quality and part of replay; exact-tier agreement remains below within-one-tier agreement, and absolute replay rates shift under substitution (Section 6).

Scope of claims. The cohort spans seven English-language generators and four length buckets. Main replay uses one source generator, one retriever family, and one decoding setting per target; each reader row is a pipeline condition tied to its source records and does not rank generators. The format and adaptation contrasts use one open-source reader, chosen for consistent performance (Section 5.2).

Framework assumptions. The 11 seed axes and phase rules are incomplete. Embedding thresholds lack universal collapse criteria. Model-authored replay items and rubrics follow a fixed schema; freezing blocks target-output adaptation but retains authoring-model dependence. Deterministic rules can under-credit faithful paraphrases; rubric-only items remain judge-sensitive. Trait and diversity audits establish neither psychological validity nor population representativeness of the personas.

Causal interpretation. The payload study uses matched contextual-stress prompts and detects exact cross-scope anchors. Paraphrased borrowing can escape that test and the deterministic packet audit, and some labeled sections fail their frozen key. Eight-persona panels are too small for stable length or episode-type estimates. The adaptation contrast covers one adapter recipe on 32 personas.

8 CONCLUSION

Many memory benchmarks evaluate a record only after it is formed. CHRONODOSSIER instead treats a long-form persona dossier as a memory representation of a known latent state, which lets construction and use be measured on the same records and traces each replay failure to a passage, a life phase, and a construction step. First, construction is part of memory: quality and collapse diverge across 311 personas, scoping each writer's payload removes detected cross-phase leakage, and staged construction improves source-item transfer. Second, retrieving the right phase does not ensure recovery: under top-three retrieval, accuracy is lower on longer dossiers and on time-dependent facts, most frontier SF failures arise in packaging or answer formation, and the length decline depends on evidence packaging. Third, for an open-source reader, re-encoding the evidence and adapting the weights move specific item families, deeper retrieval dilutes the evidence it adds, and no change reaches full-dossier accuracy. Reliable persona memory therefore depends jointly on how records are built, packaged, and read. Replay over the full cohort, more source generators, and human readers would test how far these findings extend.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks with required disclosure. Large language models are part of the method: they generate the synthetic data sets (the dossiers and cards written by the generators of Section 3 and the frozen replay items authored by GPT-5.1), answer the replay questions, annotate the adaptation data, and support the analysis as judges, with identifiers and settings in Appendix Table 12. Separately, the authors used AI assistants under their direction and review to modify manuscript text (wording, restructuring, formatting, and consistency checks) and to implement and revise analysis and release code to the authors' specifications. During drafting, they refined the language of author-written passages for clarity, concision, and consistent terminology. These assistants also reformatted run outputs into the redacted release traces and gave feedback on experimental design and the interpretation of results. We have not used generative AI tools to propose the research questions or hypotheses: the authors defined them and the study design, launched every experiment, and decided which results and claims to report. The remaining tasks with required disclosure, mathematical claims and proofs and translation, are not applicable. We have reviewed all AI-assisted work: no AI-assisted text or code entered the paper without author review, every reported number traces to a run artifact or analysis output, and the release's validation script recomputes the replay numbers from the redacted traces. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Synthetic personas can reproduce stereotypes, unrealistic correlations, and false impressions of completeness (Bender & Friedman, 2018; Gebru et al., 2021). CHRONODOSSIER provides no evidence about real populations and should not support high-stakes personalization, screening, or policy. Released artifacts need checks for memorized personal data and documented handling.

REPRODUCIBILITY STATEMENT

The anonymous repository contains the released persona cohort, construction-audit and calibration records, redacted turn-level traces for every reported replay condition, analysis outputs, and the code for construction, replay, encoding, and adaptation. A validation script recomputes the reported replay numbers from the traces. Model identifiers, decoding and retrieval settings, and seeds are recorded in Appendix Table 12, which also discloses historical configuration gaps. Replay prompts, answer keys, and completions are withheld to keep the frozen items usable as a held-out test. The repository is available at `https://anonymous.4open.science/r/ICLR_CHRONODOSSIER-4BCF`.

REFERENCES

Jaewoo Ahn, Taehyun Lee, Junyoung Lim, Jin-Hwa Kim, Sangdoo Yun, Hwaran Lee, and Gunhee Kim. TimeChara: Evaluating point-in-time character hallucination of role-playing large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3291–3325, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.197. URL https://aclanthology.org/2024.findings-acl.197/.

Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi: 10.1162/tacl_a_00041. URL https://aclanthology.org/Q18-1041/.

Jiawei Cao, Jie Ouyang, Mingyue Cheng, Zhaomeng Zhou, Chunli Liu, Yupeng Li, Zirui Liu, and Shijin Wang. Re^3: Relevance & recency retrieval for mitigating temporal hallucination. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25735–25760, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1180. URL https://aclanthology.org/2026.acl-long.1180/.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. Dense X retrieval: What retrieval granularity should we use? In *Proceedings*

of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 15159–15177, Miami, Florida, USA, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.emnlp-main.845/.

Ziyang Chen, Erxue Min, Xiang Zhao, Yunxin Li, Xin Jia, Jinzhi Liao, Jichao Li, Shuaiqiang Wang, Baotian Hu, and Dawei Yin. A question answering dataset for temporal-sensitive retrieval-augmented generation. *Scientific Data*, 12(1):1855, 2025. doi: 10.1038/s41597-025-06098-y. URL https://doi.org/10.1038/s41597-025-06098-y.

Jiayang Cheng, Dongyu Ru, Lin Qiu, Yiyang Li, Xuezhi Cao, Yangqiu Song, and Xunliang Cai. AMemGym: Interactive memory benchmarking for assistants in long-horizon conversations. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0856bc553d3e3b9827e5140d0ad3bf8d-Abstract-Conference.html.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract-Conference.html.

Zhiting Fan, Ruizhe Chen, Tianxiang Hu, Ru Peng, Zenan Huang, Haokai Xu, Yixin Chen, Jian Wu, Junbo Zhao, and Zuozhu Liu. OptimSyn: Influence-guided rubrics optimization for synthetic data generation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/b0051386fa004515c1fc3cdda8080b04-Abstract-Conference.html.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with 1,000,000,000 personas, 2024. URL https://arxiv.org/abs/2406.20094. arXiv:2406.20094.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. URL https://cacm.acm.org/research/datasheets-for-datasets/.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770. Technical report, arXiv:2607.02770.

Aaron Grattafiori et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783. Technical report, arXiv:2407.21783.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J. Taylor, and Dan Roth. Know me, respond to me: Benchmarking LLMs for dynamic user profiling and personalized responses at scale. In *Conference on Language Modeling*, 2025. URL https://arxiv.org/abs/2504.14225.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020. URL https://papers.nips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Han Luo and Guy Laban. SPASM: Stable persona-driven agent simulation for multi-turn dialogue generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 8455–8475, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.412. URL https://aclanthology.org/2026.findings-acl.412/.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5459, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1542. URL https://aclanthology.org/P19-1542/.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.747.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229, 2019. doi: 10.1145/3287560.3287596. URL https://doi.org/10.1145/3287560.3287596.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems, 2023. URL https://arxiv.org/abs/2310.08560. arXiv:2310.08560.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Xufang Luo, Hao Cheng, Dongsheng Li, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Jianfeng Gao. SeCom: On memory construction and retrieval for personalized conversational agents. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e56f394bbd4f0ec81393d767caa5a31b-Abstract-Conference.html.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. ACM, 2023. URL https://arxiv.org/abs/2304.03442.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388. Technical report, arXiv:2505.09388.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik R. Narasimhan, and Vishvak Murahari. PersonaGym: Evaluating persona agents and LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 6999–7022, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.368. URL https://aclanthology.org/2025.findings-emnlp.368/.

Wenlong Shi, Jianxun Lian, Mingqi Wu, Haiming Qin, Mingyang Zhou, Xing Xie, Naipeng Chao, and Hao Liao. PersonaArena: Dynamic simulation for evaluating and enhancing persona-level role-playing in large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 9685–9719, San Diego, California, United States, 2026.

Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.471. URL https://aclanthology.org/2026.findings-acl.471/.

Jizhou Tong and Sirui Zou. PersonaForge: Psychology-grounded dual-process architecture for personality-consistent role-playing agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 7845–7874, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.386. URL https://aclanthology.org/2026.findings-acl.386/.

Noah Wang, Z.Y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14743–14777, Bangkok, Thailand, 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.878. URL https://aclanthology.org/2024.findings-acl.878/.

Xiaoyang Wang, Hongming Zhang, Tao Ge, Wenhao Yu, Dian Yu, and Dong Yu. OpenCharacter: Training customizable role-playing llms with large-scale synthetic personas, 2025. URL https://arxiv.org/abs/2501.15427. arXiv:2501.15427.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1840–1873, Bangkok, Thailand, 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.102. URL https://aclanthology.org/2024.acl-long.102/.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3731–3741, 2019. doi: 10.18653/v1/P19-1363. URL https://aclanthology.org/P19-1363/.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations (ICLR)*, Singapore, 2025. URL https://openreview.net/forum?id=pZiyCaVuti.

Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5180–5197, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.356. URL https://aclanthology.org/2022.acl-long.356/.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. HELMET: How to evaluate long-context models effectively and thoroughly. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f5332c8273d02729730a9c24dec2135e-Abstract-Conference.html.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2204–2213, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1205. URL https://aclanthology.org/P18-1205/.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. Do LLMs recognize your preferences? evaluating personalized preference following in LLMs. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=QWunLKbBGF.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=uccHPGDlao.

Jinfeng Zhou, Yongkang Huang, Bosi Wen, Guanqun Bi, Yuxuan Chen, Pei Ke, Zhuang Chen, Xiyao Xiao, Libiao Peng, Kuntian Tang, Rongsheng Zhang, Le Zhang, Tangjie Lv, Zhipeng Hu, Hongning Wang, and Minlie Huang. CharacterBench: Benchmarking character customization of large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 26101–26110, 2025. doi: 10.1609/aaai.v39i24.34806. URL https://ojs.aaai.org/index.php/AAAI/article/view/34806.

Wangchunshu Zhou, Qifei Li, and Chenle Li. Learning to predict persona information for dialogue personalization without explicit persona description. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 2979–2991, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.186. URL https://aclanthology.org/2023.findings-acl.186/.

A Positioning Against Persona Resources

Table 7: Representative persona resources by representation scale and primary use. Units follow the source resources; medians are computed from the released profiles. The table positions research scope and does not rank native-task performance.

Resource	Representation	Primary use
PERSONA-CHAT	Static profile; five sentences	Dialogue
Persona Hub	Web-derived descriptor; one string	Data synthesis
RoleLLM / RoleBench	Character profile; median 97 words	Role-play training and evaluation
PersonaGym	Static profile; median 10 words	Dynamic agent evaluation
CharacterBench	Character profile; median 106 words	Character customization
TimeChara	Narrative point plus QA item	Point-in-time fidelity
ChronoDossier	**Year-bounded dossier; 1k–10k words**	**Construction audit and replay**

B Artifact Schemas and Join Keys

The released replay cohort contains five artifact-record families: dossier documents, compact cards, section records, phase capsules, and repair logs. The cohort manifest names each family and maps the four frontier replay rows to their stored run directories. The stable `run_id` and `persona_id` fields join persona-level records; `section_id` joins retrieved passages to section records; phase and time-scope fields link sections with phase-local evidence. Frozen replay metadata preserves item identifiers, item family, and time metadata while withholding prompts, gold answers, rubrics, and model completions from the blinded bundle. Source artifacts: `data/replay_cohort/README.md`, `data/replay_cohort/manifest.json`, and `data/cited_artifacts/artifact_manifest.md`. Table 8 lists each released family with its row unit and join keys.

C Construction Audit Details

Figure 3 separates the live gates applied during generation from the systemic audits applied to exported records. Table 9 lists the document-quality rubric shared by all judge providers.

Table 10 compares matched synthetic and Wikipedia biographies. The synthetic documents are more concentrated in both embedding views, with comparable lexical 5-gram overlap. Table 11 removes cohort size as a source of nearest-neighbor opportunity; Qwen remains the most concentrated generator under equal-size subsampling.

Table 8: Released artifact schema and join keys. `persona_id` links persona-level families; `run_id` disambiguates generator runs; `section_id`, `phase_id`, `eval_id`, and `episode_id` identify progressively finer records.

Artifact family	Row unit	Stable key / join	Contents and release scope
`persona_docs.jsonl`	accepted dossier	`run_id`; `persona_id`	Seed, timeline, outline, full dossier, length, and audit fields; all 311 released.
`persona_cards.jsonl`	persona card	`run_id`; `persona_id`	Compact values, facts, tendencies, style, and boundaries; all 311 released.
`persona_sections.jsonl`	dossier section	`run_id`, `section_id`; join `persona_id`	Heading, tags, time scope, text, and generation and audit metadata; complete for the replay cohort.
`phase_capsules.jsonl`	persona phase	`persona_id`, `phase_id`; `included_section_ids`	Phase bounds and phase-local excerpts used by temporal audits; complete for the replay cohort.
`persona_fix_log.jsonl`	repair attempt	`run_id`, `section_id`, stage, attempt	Gate failures, actions, before and after metrics, and acceptance; complete for the replay cohort.
`frozen_item_metadata.jsonl`	item definition	`eval_id`; join `run_id`, `persona_id`, `phase_id`	Family, subcategory, question shape, NLI relation, and time scope for 431 definitions; questions and gold text withheld during review.
`episode_results/*.jsonl`	model–mode episode	`episode_id`, `session_mode`	Aggregate turns, passes, history composition, retrieval counts, and usage for the four frontier rows; `traces/` covers every row. Turn text withheld during review.

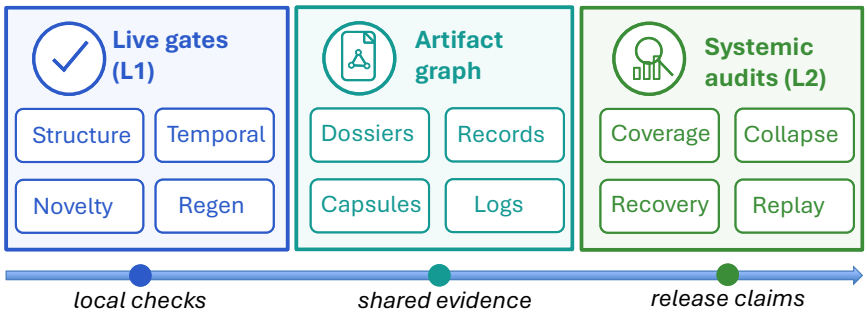


Figure 3: Two-layer audit design, expanding the audit and replay stage of Figure 1. Live gates (L1) enforce section-local admission and repair during generation; the exported artifact graph then feeds systemic audits (L2) of coverage, concentration, recoverability, and replay.

Table 9: Document-quality rubric criteria applied by every judge provider.

Criterion	Purpose
Internal consistency	Detect contradictions within and across sections.
Temporal consistency	Check adherence to the provided timeline and phase scopes.
Seed adherence	Verify faithfulness to the structured seed.
Specificity and realism	Reward concrete, plausible detail over generic prose.
Structure and readability	Evaluate organization and long-form readability.
Bias avoidance	Penalize stereotype-driven or overgeneralized content.
Retrieval readiness	Assess whether section structure supports retrieval use.
Length control	Check compliance with target length and section budgets.

Table 10: Balanced synthetic-versus-Wikipedia concentration ($n = 20$ per side). Wikipedia biographies are grounded, size-filtered, and bucket-matched by source length; the Wikipedia values are means over 21 subsets. Top1 and Pairwise are embedding similarities; 5-gram is lexical Jaccard overlap. Lower values indicate weaker concentration.

Corpus	Top1↓	Pairwise↓	5-gram↓
Synthetic	0.4039	0.2089	0.000166
Wikipedia	0.3727	0.1251	0.000210

Table 11: Equal-size ($n = 32$) concentration within each generator. Cells are medians and central 95% intervals over 10,000 subsamples of 32 documents; Gemma has exactly 32 accepted documents and therefore one possible sample. Higher similarity indicates greater concentration.

Generator	n	Top1 mean	Top1 p95
GPT-5.1	94	.547 [.523, .572]	.643 [.585, .699]
GPT-5 mini	37	.575 [.564, .584]	.642 [.626, .642]
Sonnet 4.6	35	.534 [.520, .542]	.612 [.584, .612]
Haiku 4.5	37	.519 [.506, .529]	.609 [.584, .609]
Gemma 4 26B-A4B	32	.627 [.627, .627]	.689 [.689, .689]
Llama 3.1 70B	37	.565 [.550, .576]	.664 [.639, .664]
Qwen3 32B	39	**.772 [.760, .784]**	**.838 [.821, .838]**

C.1 Phase-rubric calibration

GPT-5.1 agrees with the reference human nearly as often as the two humans agree with each other; the other two judges are harsher on disputed packets. On 25 shared phase packets, the primary human pair agrees exactly on 17/25 and within one tier on 25/25, and GPT-5.1 agrees with the reference human exactly on 16/25 and within one tier on 24/25. Gemini 3 Pro Preview and Claude Sonnet 4.5 agree exactly on 12/25 and 11/25 and within one tier on 19/25 and 20/25. Sonnet 4.5 was used because the calibration predates the main Sonnet 4.6 replay.

D Replay Configuration and Count Units

The frontier replay configuration uses target-model query planning, normalized all-MiniLM-L6-v2 embeddings, hybrid section retrieval with $k = 3$, a minimum score of 0.08, a 550-character excerpt cap, and a two-turn query-history window. The frozen cohort contains 431 unique item definitions. Scheduled retention and time-jump recurrences yield 520 rated turn instances in each reported row. These units remain separate in the main-text denominator labels. Source artifacts: `analysis/audit_and_replay/reproducibility_configuration.json`, `data/replay_cohort/README.md`, and `data/replay_cohort/manifest.json`.

The retained GPT and Sonnet frontier turn records report `retrieval_embed_model_requested` and `retrieval_embed_model_used` as `sentence-transformers/all-MiniLM-L6-v2` on rated turns. An earlier GPT run requested `Alibaba-NLP/gte-Qwen2-7B-instruct`; the released replay-cohort manifest maps

the GPT RETRIEVE row to a different run. The open-source RETRIEVE records behind Table 3 request the same MiniLM embedder but report none loaded, and the embedding score is zero for every scored section. Their retrieval therefore used lexical, heading, tag, and year matching only and returned fewer than three sections on 43 of 520 turns, none on 9. With identical heuristic queries, they retrieve the same sections as the frontier heuristic-planner control on 71 of 520 turns, while the intervention runs of Section 5.4 load the embedder, and those at $k = 3$ match that control on all 520. A later rerun of the three open-source readers with the embedder loaded, with both access modes in one run and scored with the extraction rule of Table 3, gives smaller RETRIEVE losses: 2.1, 2.3, and 6.5 points for Qwen, Llama, and Gemma, against 6.2, 8.1, and 9.6 in Table 3; no persona-bootstrap interval of the three excludes zero (`analysis/iclr/open_source_embed_rerun.json`). An attempted repeat of the Sonnet RETRIEVE run requested `Alibaba-NLP/gte-Qwen2-7B-instruct`, loaded no embedder, and retrieved no section on 54 of 520 turns; it is excluded from all reported analyses. Source artifacts: `analysis/iclr/retrieval_embedder_audit.json`, `traces/t3_gpt_retrieve.jsonl`, `traces/t3_sonnet_retrieve.jsonl`, and `data/replay_cohort/manifest.json`. Table 12 records the configuration of every generation, replay, adaptation, and retrieval component, and Table 13 gives the composition of the scripted histories.

Table 12: Recorded generation, replay, adaptation, and retrieval configuration. Provider aliases and historical local jobs did not retain resolved API revisions, model commit hashes, or a complete package lockfile; these missing pins are disclosed and not reconstructed from a later environment.

Component	Model identifier	Decoding and retrieval
GPT generation	`gpt-5.1`; provider alias, revision not logged	Minimal reasoning; provider-default sampling; stage-specific output caps.
GPT-mini generation	`gpt-5-mini`; provider alias, revision not logged	Minimal reasoning; provider-default sampling; stage-specific output caps.
Sonnet generation	`claude-sonnet-4-6`	Temperature 0 for structured calls and 0.2 for prose; stage-specific caps.
Haiku generation	`claude-haiku-4-5-20251001`	Temperature 0 for structured calls and 0.2 for prose; stage-specific caps.
Qwen generation	`Qwen/Qwen3-32B`; commit not logged	bf16; $T = 0.2$; top-$p = 0.9$; repetition penalty 1.02; at most 1,400 new tokens.
Llama generation	`meta-llama/Llama-3.1-70B-Instruct`; commit not logged	bf16; $T = 0.2$; top-$p = 0.9$; repetition penalty 1.02; at most 1,600 new tokens.
Gemma generation	`google/gemma-4-26B-A4B-it`; commit not logged	bf16; $T = 0.2$; top-$p = 0.9$; repetition penalty 1.02; at most 1,200 new tokens.
Frontier replay	`gpt-5.1`; `claude-sonnet-4-6`	GPT minimal reasoning; Sonnet $T = 0$; at most 120 output tokens; target-model query planner.
Open-source replay	Generation identifiers above	bf16 greedy decoding; repetition penalty 1.02; at most 220 output tokens; heuristic query planner; Table 3 rows without the embedder (not loaded).
Retrieval and scoring	`all-MiniLM-L6-v2`; `gpt-5.1` judge	Normalized embeddings; hybrid section retrieval; $k = 3$; score floor 0.08; 550-character excerpts; two-turn query history.
Gemma depth runs	Open-source replay model	Heuristic planner; $k \in \{3, 12\}$ under a 1,650-character total; score floor 0.08; at most 220 output and judge tokens; memoized judge calls.
Gemma adaptation	QLoRA adapter on the Gemma replay model; state chains annotated by `claude-sonnet-4-6`	4-bit base; rank 16, $\alpha = 32$, dropout 0.05 on attention projections; learning rate 5×10^{-5}; one epoch; effective batch 8; at most 3,072 tokens; seed 13.
Frontier depth runs	`gpt-5.1`; `claude-sonnet-4-6`; `gpt-5.1` judge	Heuristic planner; $k \in \{3, 12\}$ under a 1,650-character total; score floor 0.08; at most 220 output and judge tokens; memoized judge calls.

Table 13: Conversation-history composition of the frozen replay slice. Retention episodes probe supported facts, routines, and artifacts; time-jump episodes require phase selection; adversarial episodes test false-premise rejection and unsupported extrapolation. Prefilled turns are scripted history; live turns are rated target-model responses.

Episode type	Episodes	Prefilled	Live	Prefilled (%)
Retention	32	2,864	256	91.8
Time jump	32	2,744	136	95.3
Adversarial	32	2,752	128	95.6
All	96	8,360	520	94.1

E Replay Uncertainty and Answer Availability

Table 14 gives the counts and Wilson intervals behind Table 3, and Table 15 gives persona-clustered intervals for each row and each paired access contrast. Table 16 separates answer availability from answer quality for the two open-source targets affected by output extraction. Table 17 breaks frontier replay down by dossier length for the runs of Table 3 and for two packagings of heuristic-planner retrieval, which Section 5.2 compares. Figure 4 reports the frequent subcategories with the lowest frontier pass rates.

Table 14: Full replay counts behind Table 3. Each cell gives passed/rated turns, the all-turn pass rate, and the Wilson 95% half-width, $(\mathrm{hi} - \mathrm{lo})/2$, in percentage points. Nonanswers fail.

Access	Target	Overall	SF	IF	NIP
Full	GPT	479/520 (92.12 ±2.3)	339/360 (94.17 ±2.5)	88/101 (87.13 ±6.6)	52/59 (88.14 ±8.3)
Full	Sonnet	461/520 (88.65 ±2.7)	331/360 (91.94 ±2.8)	90/101 (89.11 ±6.1)	40/59 (67.80 ±11.6)
Retrieve	GPT	434/520 (83.46 ±3.2)	291/360 (80.83 ±4.1)	92/101 (91.09 ±5.7)	51/59 (86.44 ±8.8)
Retrieve	Sonnet	437/520 (84.04 ±3.1)	301/360 (83.61 ±3.8)	90/101 (89.11 ±6.1)	46/59 (77.97 ±10.4)
Full	Qwen	200/520 (38.46 ±4.2)	169/360 (46.94 ±5.1)	20/101 (19.80 ±7.7)	11/59 (18.64 ±9.8)
Retrieve	Qwen	168/520 (32.31 ±4.0)	124/360 (34.44 ±4.9)	34/101 (33.66 ±9.1)	10/59 (16.95 ±9.5)
Full	Llama	366/520 (70.38 ±3.9)	314/360 (87.22 ±3.5)	40/101 (39.60 ±9.4)	12/59 (20.34 ±10.1)
Retrieve	Llama	324/520 (62.31 ±4.2)	263/360 (73.06 ±4.6)	38/101 (37.62 ±9.3)	23/59 (38.98 ±12.1)
Full	Gemma	420/520 (80.77 ±3.4)	296/360 (82.22 ±3.9)	78/101 (77.23 ±8.1)	46/59 (77.97 ±10.4)
Retrieve	Gemma	370/520 (71.15 ±3.9)	245/360 (68.06 ±4.8)	80/101 (79.21 ±7.8)	45/59 (76.27 ±10.6)

Table 15: Length-stratified persona-clustered bootstrap intervals (95%) for all-turn replay rates and paired access-mode differences, in percent and percentage points. Each of 10,000 resamples draws personas with replacement within the four length strata and keeps all 520 paired turns of each drawn persona. Qwen and Gemma use the post-extraction scores of Table 3.

Target	Full	Retrieve	Retrieve −Full
GPT-5.1	92.1 [89.6, 94.6]	83.5 [78.5, 88.3]	−8.7 [−13.7, −3.8]
Sonnet 4.6	88.7 [85.8, 91.2]	84.0 [79.4, 88.5]	−4.6 [−10.0, +0.6]
Qwen	38.5 [31.7, 45.0]	32.3 [25.8, 38.8]	−6.2 [−13.7, +1.5]
Llama	70.4 [65.8, 74.8]	62.3 [55.6, 68.7]	−8.1 [−14.0, −2.5]
Gemma	80.8 [74.0, 86.5]	71.2 [65.0, 76.9]	−9.6 [−17.9, −1.3]

F Supplementary Replay Diagnostics

Temporal split. Table 18 gives the SF pass rates behind the temporal gap in Table 4.

Encoding format. Table 19 gives the format contrasts summarized in Section 5.4. On the 32-persona panel, the three packet rows use identical top-three sections on all 520 core turns; on the screen, both packet types select the same records.

Table 16: Answer availability for the open-source targets affected by output extraction, in percent of 520 turns. Qwen exposes an eligible answer on 273 FULL and 274 RETRIEVE turns. Every Gemma turn has an eligible leading answer, although trailing control text follows it on 500 FULL and 519 RETRIEVE turns. GPT, Sonnet, and Llama return an eligible answer on every turn.

Access	Target	All-turn rate	Rate given answer	No eligible answer
FULL	Qwen	38.46	73.26	47.50
RETRIEVE	Qwen	32.31	61.31	47.31
FULL	Gemma	80.77	80.77	0.00
RETRIEVE	Gemma	71.15	71.15	0.00

Table 17: Frontier replay by target-length bucket. Each cell gives passed turns out of the column's n rated turns, the pass rate, and the Wilson 95% half-width. The upper rows are the runs of Table 3. The lower rows use the heuristic planner and retrieve identical sections at $k = 3$; they differ in excerpt construction. Every row is a separate run with separate judge calls.

Access	Target	1k (n=136)	2.5k (n=152)	5k (n=128)	10k (n=104)
FULL	GPT	134 (98.5 ±2.4)	133 (87.5 ±5.3)	114 (89.1 ±5.4)	98 (94.2 ±4.7)
FULL	Sonnet	130 (95.6 ±3.6)	126 (82.9 ±6.0)	110 (85.9 ±6.0)	95 (91.3 ±5.5)
RETRIEVE	GPT	131 (96.3 ±3.4)	126 (82.9 ±6.0)	104 (81.2 ±6.7)	73 (70.2 ±8.7)
RETRIEVE	Sonnet	134 (98.5 ±2.4)	134 (88.2 ±5.2)	102 (79.7 ±6.9)	67 (64.4 ±9.1)
RETRIEVE, heuristic planner, 3×550 characters					
	GPT	125 (91.9 ±4.7)	120 (78.9 ±6.4)	106 (82.8 ±6.5)	70 (67.3 ±8.9)
	Sonnet	126 (92.6 ±4.5)	116 (76.3 ±6.7)	105 (82.0 ±6.6)	66 (63.5 ±9.1)
RETRIEVE, heuristic planner, $k = 3$, 1,650-character total					
	GPT	131 (96.3 ±3.4)	130 (85.5 ±5.6)	113 (88.3 ±5.6)	87 (83.7 ±7.1)
	Sonnet	131 (96.3 ±3.4)	121 (79.6 ±6.4)	113 (88.3 ±5.6)	89 (85.6 ±6.8)

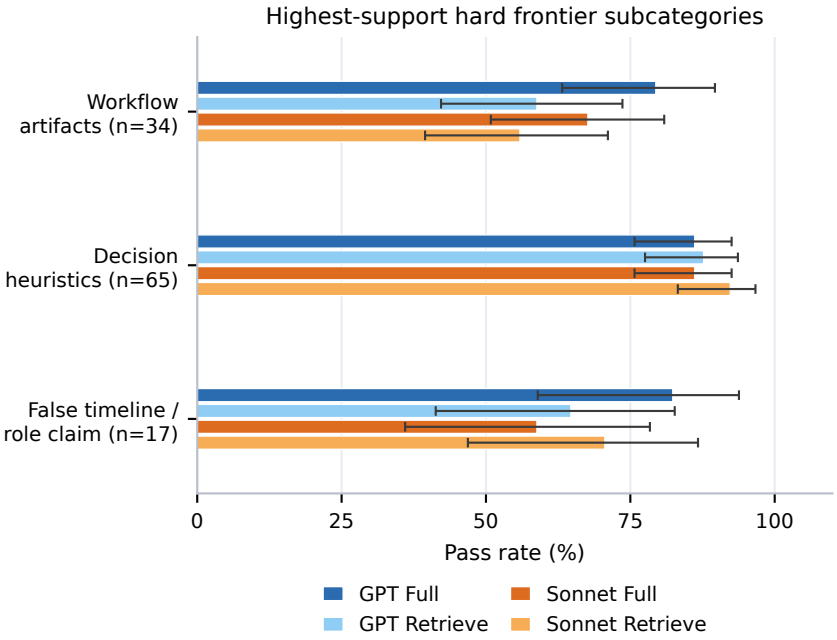


Figure 4: Frontier pass rates for the highest-support hard subcategories, with Wilson 95% intervals; support counts are rated turns. Workflow-artifact items fall under RETRIEVE for both targets, while decision-heuristic items stay above 86% in every condition.

Table 18: SF pass rates (%) over 115 phase-linked and 245 timeless items. Δ is phase-linked minus timeless in percentage points. This split was defined after the main replay analysis.

	FULL			RETRIEVE		
Target	Phase	Timeless	Δ	Phase	Timeless	Δ
GPT	95.7	93.5	+2.2	72.2	84.9	−12.7
Sonnet	89.6	93.1	−3.5	75.7	87.3	−11.7

Table 19: Recoverability by encoding format. Rows within a panel share items and retrieved evidence; the 32-persona rows come from a later run of the scoring pipeline than Table 3. IF, SF, and NIP totals are 101, 360, and 59 (upper panel) and 29, 109, and 16 (lower panel).

Panel	Encoding (Gemma, untuned)	IF	SF	NIP	Overall
32 personas (520 turns)	Full dossier (FULL)	59	335	57	451
	Raw sections, 3 × 550 characters	51	302	57	410
	Raw sections, 1,650-character total	48	298	57	403
	Versioned state records, same sections	45	309	50	404
8 personas (154 turns)	Versioned state records, 1,650 / 8,000	18 / 12	77 / 88	14 / 11	109 / 111
	Resolved-state packet, 1,650 / 8,000	16 / 18	87 / 94	12 / 10	115 / 122
	Gold sections, 8,000	17	92	10	119

Retrieval diagnostics. Frontier retrieval reaches the part of the dossier that governs each item. Over 208 unique frontier items, gold phase hit@3 is 100% and section hit@3 is 94.7% for GPT and 97.3% for Sonnet; over the 520 rated turns, phase-pure retrieval is 94.4% and 93.1%. Source artifact: `analysis/audit_and_replay/retrieval_phase_hits_publication_rag.json`.

Planner substitution. Frontier RETRIEVE results do not require target-authored queries. A fixed heuristic planner keeps phase hit@3 at 100% but lowers section retrieval; GPT moves from 434 to 421 of 520 turns (planned-minus-heuristic $+2.5$ points, 95% CI $[-2.3, +7.4]$) and Sonnet from 437 to 413 ($+4.6$ points, 95% CI $[-0.3, +9.0]$). Both intervals include zero; no planner effect is established.

Failure attribution by reader. With identical packets, every reader rarely fails when the packet satisfies the key, and the readers differ mainly on turns whose packet does not. Table 20 gives the SF failure attribution of Section 6 for the two frontier RETRIEVE rows of Table 3 and for the depth contrast, in which GPT-5.1, Sonnet 4.6, and Gemma receive byte-identical packets on all 520 turns at each depth; answer-formation failures stay between 2 and 10 per reader. At $k = 12$, retrieval failures vanish and packaging failures rise for every reader, the dilution reported in Section 5.4. Source artifact: `analysis/iclr/sf_packet_attribution_extended.json`.

Gold-section and fixed-budget design. Items without section labels, mainly prompt-injection refusals and card-grounded items, retain the compact card in the gold-section condition and are reported separately; the condition controls evidence selection but retains the bounded packet and downstream answer process. The retrieval-budget study holds the query planner, answer prompt, conversation history, target model, and scorer fixed while varying $k \in \{1, 3, 5\}$. Under the *per-passage* policy, each passage receives up to 550 characters, and evidence volume grows with k; under the *fixed-total* policy, at most 1,650 characters are distributed across selected passages. The gold-section condition uses the same 1,650-character cap. Every turn records selected section IDs, gold hit and recall, realized evidence characters, per-section allocation, the effective dense embedder, tokens, and cost. The fixed-total diagnostic uses eight personas, two from each length bucket, and 130 rated turns per condition; it tests the intervention mechanics and does not replace the 32-persona, length-stratified comparison.

Fixed-budget evidence diagnostic. Perfect section selection adds little on this panel: under the same packaging, gold sections add zero net passes over $k = 3$ for GPT-5.1 and three for Sonnet 4.6

Table 20: SF failure attribution on the 360 SF turns of the 520-turn core. A failure is assigned to retrieval (no labeled section selected), packaging (a selected labeled section satisfies the frozen key and the packet does not), or answer formation (the packet satisfies the key); "No key" counts failures whose labeled sections do not satisfy the key. The last column counts passes without detected packet support.

Evidence	Reader	Pass	Fail	Retr.	Pack.	Answer	No key	Pass, no support
Table 3 RETRIEVE	GPT-5.1	291	69	3	29	23	14	30
(model planner, 3×550)	Sonnet 4.6	301	59	0	34	16	9	28
Identical packets	GPT-5.1	315	45	2	28	4	11	102
($k = 3$, 1,650 total)	Sonnet 4.6	314	46	1	29	4	12	101
	Gemma	298	62	3	43	6	10	87
Identical packets	GPT-5.1	312	48	0	40	2	6	87
($k = 12$, 1,650 total)	Sonnet 4.6	302	58	0	50	6	2	81
	Gemma	295	65	0	47	10	8	78

(Table 21). Both targets pass over 94% of turns with gold sections, leaving little room for a retrieval gain on this panel. The counts bound the observed return to retrieval improvement; they do not measure answer quality.

Table 21: Eight-persona fixed-total diagnostic: answer passes and mean labeled-section recall.

Target	Gold	$k = 1$	$k = 3$	$k = 5$	Recall@3	Recall@5
GPT-5.1	123/130	115/130	123/130	122/130	76.2%	88.2%
Sonnet 4.6	124/130	111/130	121/130	121/130	80.1%	89.3%

Top-3 scores highest on this panel. Moving from $k = 1$ to $k = 3$ adds 8 GPT and 10 Sonnet passes (paired wins/losses 11/3 and 11/1). Moving to $k = 5$ raises mean labeled-section recall by 12.0 and 9.2 points without an answer gain, consistent with dilution under the fixed total: across 93 SF turns, labeled-section selections rise slightly (88 to 91 for GPT, 91 to 93 for Sonnet) while key-satisfying packets fall (80 to 76 and 80 to 77). These are not population estimates, and the packet audit covers SF only; IF and NIP lack sufficiency labels. Source artifacts: `analysis/iclr/experiment_summary_gpt51.json`, `analysis/iclr/experiment_summary_sonnet46.json`, and the `traces/diagnostic_*` files.

G PROMPTS, RUBRICS, EXCLUSIONS, AND RELEASED TABLES

Replay prompt construction, query planning, evidence selection, fixed-total allocation, response extraction, and scoring are implemented in `code/replay/eval_episode_runner.py`. The gold-section and retrieval-budget conditions are defined in `code/replay/run_iclr_retrieval_experiments.py`, which also keeps the earlier weighted metadata matcher for replication; aggregate comparisons are produced by `code/replay/analyze_iclr_retrieval_experiments.py`. SF packet attribution is produced by `code/replay/analyze_sf_packet_attribution.py`. Phase-evaluation rubrics and human instructions are indexed by `data/cited_artifacts/artifact_manifest.md`. Redacted turn-level traces for every reported replay condition are in `traces/`, and `code/analysis/iclr/recompute_from_traces.py` recomputes the reported replay numbers from them.

The replay-cohort README records the blinded exclusions and explains why 431 frozen definitions produce 520 rated turns. Per-row episode aggregates and analysis outputs retain family counts, clustered uncertainty, temporal strata, answer availability, judge substitution, retrieval diagnostics, and SF packet attribution. Source artifacts, under `analysis/audit_and_replay/`: `replay_clustered_uncertainty_summary.json`, `replay_temporal_stratified_summary.json`, `open_source_output_sensitivity_summary.json`, `judge_sensitivity_summary.json`, `retrieval_phase_hits_publication_rag.json`, and `sf_packet_attribution_summary.json`; see also `data/replay_cohort/README.md`.

H INTERVENTION DIAGNOSTICS

Gold-section evidence. The gold-section condition builds each packet from the exact labeled section identifiers, after a check that every identifier resolves to a stored section, and enforces the same total evidence cap as the fixed-budget RETRIEVE conditions. The eight-persona diagnostic covers both frontier targets, with 130 rated turns per condition. Gold-section evidence is evaluated only on eight-persona panels: this diagnostic and the 8,000-character screen of the encoding-format comparison (Appendix F). Source artifacts: `code/replay/run_iclr_retrieval_experiments.py`, `code/replay/eval_episode_runner.py`, and the `traces/diagnostic_*` files.

Output delimitation. The contrasts of Section 5.4 score each Gemma response in full, including text generated after its end-of-turn marker; Table 3 removes that text. Removing it leaves the adaptation IF loss (-14.9 points, $p = 0.011$) and the encoding NIP loss (-13.6, $p = 0.008$), lowers overall accuracy under versioned state records by 4.0 points ($p = 0.038$), and reduces the depth IF gain to $+5.9$ ($p = 0.24$). Source artifact: `analysis/iclr/gemma_interventions_postextraction.json`.

Parameter-efficient adaptation diagnostics. The earlier adaptation stages show reliable format and citation-control effects and no answer-accuracy gain; the main adaptation contrast in Section 5.4 uses the SFT adapter from these experiments. An initial extractive QLoRA pilot on Gemma 4 changed matched replay from 102/130 to 105/130 (7 adapted-only and 4 base-only passes), and under the same chain evidence, a later state-chain adapter passed 134/154 against 129/154 for the SFT adapter; both changes are inconclusive by exact McNemar ($p = 0.549$ and $p = 0.227$). The state-chain adapter raised exact output-schema compliance from 30/154 to 151/154. Sparse-active refinement then reduced overcitation from 79/128 to 45/128 defined-target turns while overall answer accuracy stayed at 104/154 and minimal-support satisfaction changed from 109/128 to 108/128. Across four correction-stage seeds, per-turn citation precision rose from 69.8% untuned to 74.0–76.8%. Source artifacts: `analysis/iclr/sparse_active_screen_analysis.json`, `analysis/iclr/seed_noise_floor_analysis.json`, and the `traces/screen_*` files; the analyses of the extractive and state-chain pilots are not in the release.