ClinTraceBench: Source-Verifiable Longitudinal Clinical Reasoning over EHR-Derived Dialogues

Anonymous ACL submission

Abstract

Clinical LLM assistants must reason over multi-visit patient trajectories, yet whether the compact history representations used to scale them—retrieval, structured timelines, LLM summaries, agentic memory—preserve the longitudinal signal clinical reasoning needs has not been measured. We introduce ClinTraceBench: 385 MIMIC-IV-derived verified dialogues with event-ID provenance, a nine-task taxonomy (T1–T9), and L0–L4 deterministic + L5 human-audit validation (98.92% agreement). We evaluate eight history representation strategies (*no-context-blind*, *last-visit-only*, *full-context*, *dense-retrieval* with BGE-M3, *structured-timeline*, *llm-summary*, *Mem0*, *A-Mem*) across four backbones (DeepSeek-V3, GPT-4o-mini, Haiku 4.5, Sonnet 4.6) on 6,271 questions: 32 cells, 200,672 predictions. Four findings: (SP4) a controlled T3 injection reveals deterministic preservation failure—four compressed representations collapse to 50.0% on a balanced 300/300 split (the constant-No floor), independent of model capacity; (SP1) compressed strategies pay an aggregation tax on multi-visit trends and cross-patient comparisons; (SP2) the blind-to-full gap spans +29.8 pp (GPT-4o-mini) to +62.7 pp (Haiku); (SP3) abstention scales non-monotonically with context length. On the Pareto frontier, Haiku dominates Sonnet under *full-context* ($25.76 vs. $106.21, 0.706 vs. 0.698), inverting the "biggest backbone wins" heuristic. We will release the benchmark, predictions, and the injection generator.

1 Introduction

Clinical assistants built on LLMs must reason over long, multi-visit patient trajectories: retrieving labs, tracking trends across encounters, linking findings to problems within a visit, recalling treatment, comparing patients, and abstaining when the record is silent. As context windows grow (Liu et al., 2024; Bai et al., 2024), feeding the full chart is increasingly tractable but expensive, latency-bound, and—as we show—interacts non-monotonically with abstention. Practitioners therefore rely on compact history representations: retrieval-augmented generation (Lewis et al., 2020; Chen et al., 2024), structured timelines, LLM-generated summaries, and agentic memory systems such as *Mem0* (Chhikara et al., 2025) and *A-Mem* (Xu et al., 2026). Whether these representations preserve enough signal for longitudinal clinical reasoning is the empirical question this paper addresses.

Three gaps make this question hard to answer with existing resources. First, EHR-derived clinical benchmarks (Kweon et al., 2024; Fleming et al., 2024; Ma et al., 2024; Ben Abacha et al., 2024) are dominated by single-encounter or single-document prompts; they do not stress multi-visit aggregation, encounter-local linkage, or abstention under unstated facts. Second, open-domain memory benchmarks (Maharana et al., 2024; Wu et al., 2025; Hu et al., 2025) evaluate dialogue memory but lack clinical semantics and event-level source provenance, so a failure cannot be traced to a specific dialogue turn. Third, when frontier LLMs are evaluated on long clinical context, known long-context pathologies (lost-in-the-middle attention (Liu et al., 2024), miscalibrated abstention (Kadavath et al., 2022)) have not been characterized at the level of specific clinical cognitive operations.

ClinTraceBench fills these gaps with 385 MIMIC-IV-derived verified dialogues (Johnson et al., 2023) carrying event-ID provenance, a nine-task taxonomy covering fact recall, temporal trends, encounter-level linkage, retrospective propagation, contradiction, cross-patient comparison, treatment recall, observed treatment response, and abstention, and L0–L4 deterministic + L5 human-audit validation [1] (98.92%). We evaluate eight rep-

[1] L0–L5 denote six validation gates applied to each (anchor, gold) pair: L0 schema check, L1 source verification, L2 gold consistency, L3 deduplication, L4 stratification balance, and

resentation strategies against four frontier backbones (DeepSeek-V3, GPT-4o-mini, Haiku 4.5, Sonnet 4.6) on a fixed 6,271-question set: 32 cells, 200,672 predictions, $527.85. Contributions:

1. **Benchmark construction.** A reproducible pipeline turning real MIMIC-IV trajectories into verified, source-anchored dialogues with L0–L4 deterministic + L5 human spot-check validation—linking every gold answer to a specific dialogue turn, a property absent from prior memory benchmarks (Maharana et al., 2024; Wu et al., 2025).
2. **Task taxonomy.** Nine tasks covering longitudinal axes existing clinical and memory benchmarks miss, organized by the cognitive operation a clinician performs over a chart (T1–T9, defined in §3).
3. **Controlled preservation probe.** A sentence-level injection probe (T3) that isolates whether a representation preserves a clinically relevant signal independent of backbone capacity—a methodology generalizable beyond clinical settings.
4. **Findings.** Full context wins on pooled accuracy but is not uniformly best; dense retrieval is competitive at lower cost; compressed and agentic memory representations exhibit task-specific failures (aggregation tax, sentence-signal loss, abstention drift); the cost–quality Pareto frontier inverts the "biggest backbone wins" heuristic.

2 Related work

EHR-derived clinical evaluation. EHRNOTEQA (Kweon et al., 2024) curates MIMIC-IV discharge-summary QA pairs validated by clinicians; MEDALIGN (Fleming et al., 2024) similarly grounds clinician-written instructions in real EHR notes; CLIBENCH (Ma et al., 2024) extends MIMIC-IV to multi-decision clinical tasks (diagnosis, procedure, lab, prescription); MEDEC (Ben Abacha et al., 2024) probes medical error detection on clinical notes; DR.BENCH (Gao et al., 2023) targets progress-note diagnostic reasoning over multi-encounter MIMIC-III records; MEDHELM (Bedi et al., 2025) provides a multi-capability evaluation framework for medical LLMs. These benchmarks evaluate models against real chart data, but each is dominated by single-document or single-encounter prompts. None stress multi-visit aggregation, encounter-local attribution, or abstention under unstated facts.

Temporal and longitudinal clinical reasoning. The i2b2 temporal challenges (Sun et al., 2013) established temporal-relation extraction on clinical notes; TIMER (Cui et al., 2025) introduces temporal instruction tuning over multi-encounter EHR data; recent work (Lal et al., 2024) evaluates frontier LLM temporal reasoning on synthetic clinical timelines. None of these provide event-level source-provenance from the answer back to a specific dialogue turn, nor a controlled probe that isolates signal-preservation failure modes from backbone reasoning capacity.

Long-context and retrieval evaluation in NLP. Long-context benchmarks evaluate retrieval and reasoning across extended inputs: LONGBENCH (Bai et al., 2024) and needle-in-a-haystack probes (Kamradt, 2023) characterize attention drift; Liu et al. (2024) document the lost-in-the-middle phenomenon. Retrieval-augmented generation (Lewis et al., 2020) and *dense-retrieval* embeddings (Chen et al., 2024) form the foundation of one strategy family we evaluate. Kadavath et al. (2022) establish calibration and abstention as a measurable LLM capability, which our T9 task operationalizes in a clinical setting.

Memory benchmarks and agentic memory. LOCOMO (Maharana et al., 2024) evaluates multi-session conversational memory; LONGMEMEVAL (Wu et al., 2025) extends to longer, more diverse memory horizons; MEMORYAGENTBENCH (Hu et al., 2025) evaluates incremental multi-turn memory; a recent survey (Zhang et al., 2025) reviews memory mechanisms in LLM agents. The open-domain memory lineage motivates agentic-memory methods we evaluate Mem0 (Chhikara et al., 2025), AMEM (Xu et al., 2025), MemGPT (Packer et al., 2023). We repurpose this lineage to ask whether memory-style representations preserve enough longitudinal clinical signal to compete with *full-context* feeding on tasks designed around real EHR trajectories.

Position of our work. We are not the first MIMIC-IV (Johnson et al., 2023) benchmark, not the first temporal benchmark, and not the first memory benchmark. Our contribution is the integration: real EHR-derived longitudinal trajectories, verified dialogue substrate with event-level provenance, a nine-task taxonomy covering fact / temporal / linkage / retrospective / contradiction / comparison / treatment / response / absten-

L5 human audit. Full criteria in Appendix A.

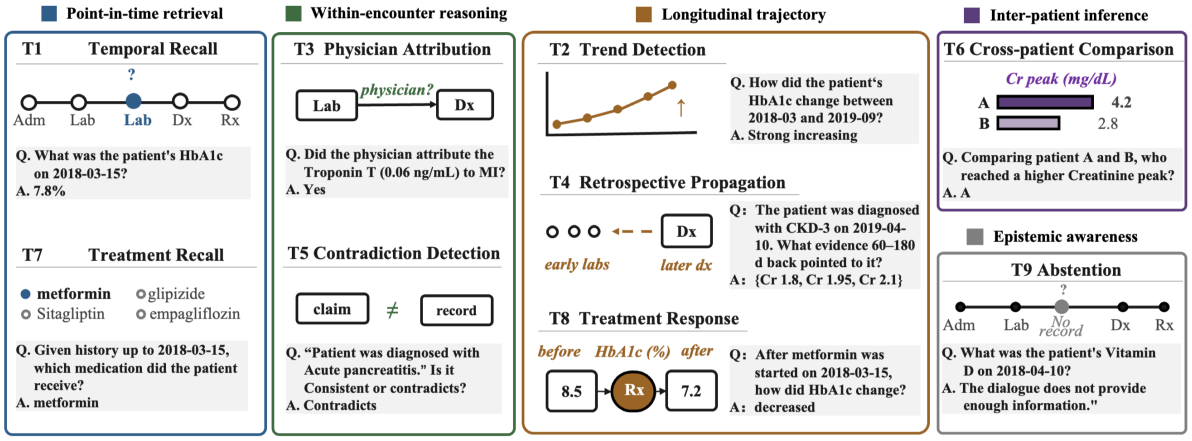

Figure 1: ClinTraceBench task taxonomy. The nine tasks (T1–T9) are grouped by five cognitive operations: point-in-time retrieval (T1, T7), within-encounter reasoning (T3, T5), longitudinal trajectory (T2, T4, T8), inter-patient comparison (T6), and epistemic awareness (T9). Each card shows the task identifier, a schematic of the source anchor, and one example question (Q) with its deterministic gold answer (A). Per-task question counts are broken out in Table 1.

tion, L0–L4 deterministic + L5 human-audit validation, and a controlled-injection probe that isolates representation-level signal preservation from backbone capacity—enabling head-to-head paired comparison of history representation strategies.

3 Benchmark design

Figure 1 shows the T1–T9 task taxonomy organized by the cognitive operation a clinician performs when reading a longitudinal record; Figure 2 shows the source-verifiable construction pipeline, whose per-stage Output bands report the full-cohort headline counts (400 patients yielding 385 verified dialogues, 6,271 evaluation questions, 200,672 predictions) and per-stage validation gates.

Cohort. We sample 400 MIMIC-IV (hospital-table) patients, balanced across four target diseases (diabetes, hypertension, CKD, CAD; 100 each) and stratified within each disease into low / medium / high complexity tertiles (33/34/33) by event, admission, span, and medication counts. Inclusion requires $\geq$ 2 admissions $>$ 7 days apart, $\geq$ 5 labs, $\geq$ 2 prescriptions, and $\geq$ 2 ICD codes in the index disease; sampling is deterministic. 15 patients fail dialogue invariants during synthesis and are dropped, yielding 385 verified dialogues. Lab extraction uses a fixed 18-lab panel with canonical units and clinical-priority deduplication (full schema in Appendix). Each record is transformed into a multi-visit patient–clinician dialogue, preserving structured fields (dates, lab values, ICD / procedure codes, medications) inside narrative turns. Consistent with prior evidence that frontier LLMs encode clinical knowledge (Singhal et al., 2023), we use a held-out frontier generator with deterministic templates for structured fields. MIMIC-IV is HIPAA-compliant and we introduce no additional identifiers.

Task suite. The nine tasks are:

- **T1 — Numeric lab retrieval.** Given a lab name and date, return the value.
- **T2 — Trend classification.** Direction (increased / decreased / stable) and magnitude across two-to-five visits.
- **T3 — Controlled attribution detection.** For selected same-encounter finding–problem pairs, the dialogue either contains or omits a physician-attribution sentence of the form "physician noted: finding is consistent with problem" (balanced yes / no). Four strategies whose representation is built upstream of injection (Mem0, A-Mem, *llm-summary*, *structured-timeline*) cannot recover the sentence by construction; T3 is therefore a probe of representation-level signal preservation, not an apples-to-apples accuracy comparison (Section 5.5).
- **T4 — Retrospective propagation.** Given a later diagnosis, list the earlier findings constituting its evidence. Scored as binarized set-F1 on event UIDs (1 iff exact match), so T4 pools identically with accuracy. Cross-visit and longitudinal.
- **T5 — Controlled contradiction detection (exploratory).** Does the chart contain a controlled self-contradictory statement? Contradicts-skewed by construction (455 / 40), giving a trivial majority baseline of 0.919 that no cell beats; re-

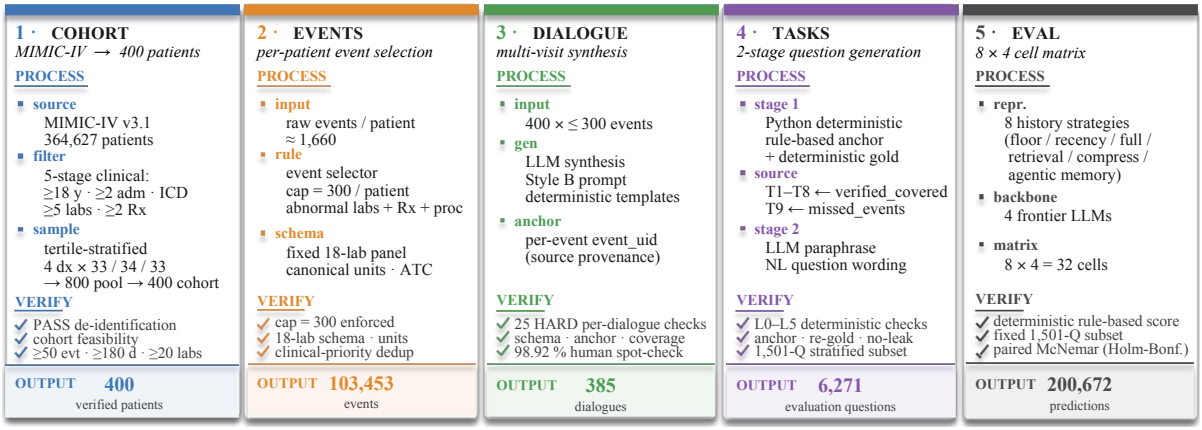

Figure 2: ClinTraceBench construction pipeline. 1. Cohort: MIMIC-IV 364,627 → 400, 4-disease balanced + complexity-tertile stratified. 2. Events: fixed 18-lab panel, deterministic event IDs for provenance. 3. Dialogue: Sonnet 4.6 synthesis + DeepSeek/Qwen3-Max rescue; verifier v3.2 (10 HARD + 3 SOFT); 385/400 pass. 4. Tasks: Python-generated anchor+gold, Sonnet paraphrases to questions; L0–L4 deterministic + L5 human audit yields the 6,271-question set (98.92% agreement). 5. Evaluation: $8 \times 4 = 32$ cells, 200,672 predictions. Each panel lists its PROCESS, VERIFY, OUTPUT.

ported only for relative ordering.

- **T6 — Cross-patient comparison.** Three subtypes T6a / T6b / T6c (number of diagnoses, max lab value, delta value).
- **T7 — Treatment fact recall.** Four-option clinical-management question grounded in documented treatment events; pool split between medication and procedure subtypes to avoid within-task class bias.
- **T8 — Observed treatment response (exploratory).** Did a lab increase, decrease, or stay similar after a documented treatment event? Exploratory: concurrent-medication annotations were not retained, so T8 tracks post-treatment lab change, not causal response.
- **T9 — Abstention.** All gold answers are "insufficient information"; T9 accuracy equals the abstention rate, with over-answer rate = 1 − abstention. A trivial always-abstain classifier scores 100% by construction, so T9 is used as a relative ordering, not an absolute benchmark.

Evaluation set. The L0–L4 + L5 pipeline yields a 6,271-question stratified set (T1 = 1500, T2 = 889, T3 = 600, T4 = 53, T5 = 495, T6 = 700, T7 = 400, T8 = 134, T9 = 1500), fixed across all 32 cells to support paired McNemar tests. At $p = 0.5$, the pooled Wilson 95% CI is ±0.6 pp; per-task CIs range from ±1.3 pp (T1, T9) to ±9 pp (T4); per-task minimum-detectable effects span 5.1–27.2 pp. Pairwise headline comparisons use Holm–Bonferroni at $\alpha = 0.05$ ($m = 12$); all 12 remain significant after correction.

Headline metric. We report *pooled* accuracy (over all 6,271 questions) and *macro* accuracy (mean of the nine per-task accuracies). Because per-task sizes are unbalanced ($n_{T1} = n_{T9} = 1500$ vs. $n_{T4} = 53$), pooled accuracy is T1/T9-dominated; macro accuracy and the per-task surface (Table 1) should be consulted for representation comparisons that should not be task-mix-weighted.

4 History representation strategies and backbones

History representation strategies. Eight strategies span the principal context-access choices, grouped into six families:

- **Floor.** *no-context-blind* — the model sees only the question.
- **Recency.** *last-visit-only* — only the most recent encounter.
- **Full.** *full-context* — entire multi-visit dialogue fed verbatim; upper bound for context coverage.
- **Retrieval.** *dense-retrieval* — BGE-M3 chunk embeddings per patient, top-K=5.
- **Compression.** *structured-timeline* (per-encounter table) and *llm-summary* (single 500-token DeepSeek summary).
- **Agentic memory.** Mem0 (extracted facts, DeepSeek-prepared, top-K=5) and A-Mem (Zettelkasten-style atomic notes, DeepSeek-prepared).

Backbones. Inference is routed through OpenRouter: DeepSeek-V3 (deepseek-chat-v3.1, 64k), GPT-4o-mini (gpt-4o-mini, 128k), Claude Haiku 4.5 (claude-haiku-4.5, 200k), and Claude Sonnet 4.6 (claude-sonnet-4.6, 200k). Prompt-prefix caching is enabled across all

Table 1: HealthMemBench main results: 8 strategies × 4 backbones × 9 tasks. Per-cell accuracy (%) for every (strategy, backbone, task) triple. Bottom rows: Pool. = accuracy pooled across tasks weighted by subset size; Macro = equal-task-weight mean; Δ_{blind} = Macro minus the same-backbone *no-context-blind* Macro (pp). Backbones: DS = DeepSeek-V3, GPT = GPT-4o-mini, Hk = Claude Haiku 4.5, Sn = Claude Sonnet 4.6. Strategies: blind = *no-context-blind*; last-vis = *last-visit-only*; full = *full-context*; dense = *dense-retrieval*; struct. = *structured-timeline*; summ. = *llm-summary*. Blue heat: accuracy (<30 30–45 45–60 60–70 ≥ 70). Green heat (Δ_{blind} row only): (10–20 20–40 ≥ 40).

	blind				**last-vis**				**full**				**dense**				**struct.**				**summ.**				**Mem0**				**A-MEM**			
Task	DS	GPT	Hk	Sn	DS	GPT	Hk	Sn	DS	GPT	Hk	Sn	DS	GPT	Hk	Sn	DS	GPT	Hk	Sn	DS	GPT	Hk	Sn	DS	GPT	Hk	Sn	DS	GPT	Hk	Sn
T1	0	0	1	0	16	15	13	12	81	78	81	84	80	74	78	79	89	89	92	92	36	36	35	36	52	52	51	49	49	42	43	49
T2	17	33	0	30	33	35	18	35	38	26	35	47	35	27	43	43	30	21	29	41	26	33	32	34	33	30	35	39	33	32	37	38
T3	50	49	50	50	56	54	55	51	94	93	87	89	83	85	87	77	51	51	50	50	50	49	50	50	49	48	50	50	50	48	48	50
T4	13	8	19	11	6	11	17	11	23	28	19	25	21	23	15	6	6	21	25	17	15	17	19	11	4	13	9	15	8	11	9	13
T5	25	39	17	38	79	58	37	64	86	90	85	84	87	89	86	84	75	78	87	86	83	67	64	63	80	73	72	75	78	69	74	78
T6	36	31	0	10	42	47	47	42	54	50	51	52	47	39	44	46	61	53	52	56	44	39	42	42	46	42	45	43	50	46	44	48
T7	26	24	1	20	36	30	14	37	94	87	97	95	90	86	94	93	64	62	51	69	51	48	35	45	73	64	54	66	65	62	44	62
T8	23	25	17	26	31	35	18	23	44	27	46	47	39	28	40	40	45	40	54	49	23	25	19	22	37	36	31	30	37	34	29	37
T9	36	69	0	20	88	90	99	93	70	70	91	75	77	71	92	79	56	66	48	44	80	78	94	74	81	82	94	73	82	78	97	73
Pool.	24	33	8	20	47	46	42	45	67	63	71	70	66	60	70	66	59	59	58	60	48	47	50	46	55	54	56	52	55	51	54	53
Macro	25	31	12	23	43	42	35	41	65	61	66	67	62	58	64	61	53	53	54	56	45	44	43	42	51	49	49	49	50	47	47	50
Δ_{blind}	—	—	—	—	18	11	24	18	40	30	54	44	37	27	53	38	28	22	43	33	20	13	32	19	26	18	37	26	25	16	36	27

backbones (90.7 M cache-read tokens, 17.1% of input). Provider caching settings, T8 / T9 output-token caps, and the T9 abstention parser are in Appendix D.

Memory prep model. Both agentic-memory strategies extract artifacts (facts or atomic notes) from the dialogue once before any question is asked. We hold the prep model fixed at DeepSeek-V3 across all four answering backbones, isolating answering-side use of memory from construction-side extraction. A consequence: non-DeepSeek Mem0 and A-Mem cells run a heterogeneous pipeline (DeepSeek-extracted artifacts answered by GPT-4o-mini / Haiku / Sonnet); we discuss the SP2 impact in §Limitations.

5 Results

5.1 Main accuracy

Table 1 reports the full strategy-×-backbone-×-task cube; the three right-margin columns give pooled, macro, and Δ_{blind} (macro lift over the same-backbone blind). *full-context* is best on every backbone (pooled 67.2 / 63.0 / 70.6 / 69.8 for DeepSeek / GPT-4o-mini / Haiku / Sonnet). *dense-retrieval* follows on every backbone (60.4–69.7), within 1–4 pp of *full-context*—BGE-M3 with $K=5$ captures most but not all of what *full-context* provides. *structured-timeline* sits third (58.4–59.6) and is strikingly backbone-invariant (± 0.6 pp). *Mem0* and *A-Mem* cluster at 50.8–55.9; *llm-summary* and *last-visit-only* trail at 41.6–49.5. Macro accuracy preserves the same ordering as pooled. The per-task surface exposes the difficulty structure: even under *full-context*, T2 / T4 / T6 / T8 / T9 remain hard while T1 / T3 / T5 / T7 are tractable; compressed strategies show the T3 constant-No collapse (the band of 50.0 cells, Section 5.5) and the aggregation tax on T2 / T6 / T8.

Table 2 decomposes the context lift (vs. same-backbone blind) for three regimes (*full-context*, *dense-retrieval*, and the agentic-memory mean of *Mem0* and *A-Mem*); blue shades positive lift, red the rare cells where context hurts.

Table 2: Per-cell context-lift over *no-context-blind*. Accuracy differences (pp) between three context regimes (*full-context*, *dense-retrieval*, mem-avg = mean of Mem0 and A-MEM) and the same-backbone, same-task *no-context-blind* baseline. Macro = mean over T1–T9. Diverging color encodes lift (≤ -5 -5 to 0 0–5 5–15 15–30 30–60 ≥ 60). See Appendix H for heatmap visualization.

Strategy	**BB**	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**	**Macro**
full	DS	+81	+20	+44	+9	+61	+19	+69	+21	+35	+40
	GPT	+78	-8	+44	+21	+51	+19	+63	+2	+1	+30
	Hk	+81	+35	+37	0	+68	+51	+95	+29	+91	+54
	Sn	+84	+17	+39	+13	+46	+42	+76	+21	+54	+44
dense	DS	+80	+18	+33	+8	+62	+11	+64	+16	+42	+37
	GPT	+74	-7	+35	+15	+50	+8	+62	+2	+3	+27
	Hk	+77	+43	+37	-4	+69	+44	+93	+23	+92	+53
	Sn	+78	+13	+27	-6	+46	+36	+73	+13	+59	+38
mem avg	DS	+50	+16	0	-8	+54	+12	+44	+14	+46	+25
	GPT	+47	-2	-2	+5	+32	+13	+38	+9	+12	+17
	Hk	+46	+36	-1	-9	+56	+45	+48	+13	+95	+36
	Sn	+48	+8	0	+3	+38	+35	+44	+8	+53	+26

5.2 SP1: The aggregation bottleneck

We index findings along two complementary axes: SP1 (task × strategy, this subsection) and SP2 (backbone × strategy, Section 5.3). The two are entangled—GPT-4o-mini's -7.8 pp on T2 with summary memory is both an SP1 observation (T2 demands aggregation; summaries destroy it) and an SP2 one (GPT-4o-mini has the narrowest blind-to-full gap).

Compressed representations lose to *full-context* because preprocessing discards signal needed at answer time. The penalty is sharpest on multi-value aggregation—T2 (trend across visits), T6b / T6c (max and delta across patients), and exploratory T8. On T8, *full-context* reaches 0.269–0.470 across backbones while *Mem0*, *A-Mem*, and *llm-summary* fall to 0.291–0.373, 0.291–0.373, and 0.194–0.254 respectively. T7 makes the contrast cleanest by rewarding single-fact recall over aggregation: *full-context* reaches 0.942–0.965 (DeepSeek / Haiku / Sonnet), while *Mem0* drops to 0.535–0.733 and *llm-summary* to 0.349–0.512. Figure 3 visualizes the T2 / T6 bottleneck. The mechanism is structural: agentic-memory representations never store per-visit value pairs, and a 500-token *llm-summary* cannot encode a multi-visit trajectory. Dense retrieval mitigates this—original text remains addressable—but still fails when aggregation crosses non-contiguous chunks.

5.3 SP2: Backbone-specific extraction profiles

The blind-to-full gap diagnoses how much of the answer each backbone can extract from the chart when given full access:

- **Haiku**: $0.079 \rightarrow 0.706$, gap $+62.7$ pp (largest).
- **Sonnet**: $0.203 \rightarrow 0.698$, gap $+49.6$ pp.
- **DeepSeek-V3**: $0.236 \rightarrow 0.672$, gap $+43.6$ pp.
- **GPT-4o-mini**: $0.332 \rightarrow 0.630$, gap $+29.8$ pp (smallest).

Haiku exhibits the cleanest context utilization: lowest blind (0.079), tied top under *full-context*, largest 62.7 pp lift. Sonnet and DeepSeek-V3 occupy the middle. GPT-4o-mini is the outlier: highest blind (0.332), lowest *full-context* (0.630), narrowest lift—a pattern consistent with a hallucination floor that subsequent SP3 evidence on T9 reinforces. "How much a model uses given context" therefore varies more across backbones than overall accuracy suggests, and backbone choice cannot be reduced to a single capacity dimension.

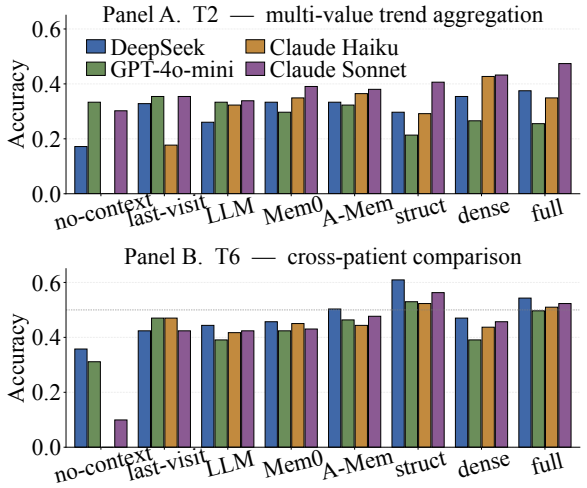

Figure 3: Accuracy on two longitudinal-reasoning bottlenecks, grouped by strategy (x-axis) and backbone (color). (A) T2 trend aggregation (192 q/cell). (B) T6 cross-patient comparison (151 q/cell, subtypes T6a/T6b/T6c). Dotted line in Panel B = 0.5 reference (suppressed in A where all bars sit below 0.5). Missing Haiku no-context and Sonnet's near-zero on the same column are parser-strictness artifacts on verbose abstentions. Even *full-context* stays well below ceiling: T2/T6 are reasoning bottlenecks distinct from point-fact recall.

5.4 SP3: Non-monotonic abstention under longer context

T9 probes abstention over unstated facts; all 1,500 gold answers are "insufficient information", so accuracy equals the abstention rate and over-answer rate $= 1 -$ abstention. Per-cell:

- *no-context-blind*: abstention 0.355 / 0.685 / 0.000 / 0.204 (DeepSeek / GPT-4o-mini / Haiku / Sonnet); over-answer 0.645 / 0.315 / 1.000 / 0.796.
- *full-context*: abstention 0.701 / 0.698 / 0.907 / 0.747; over-answer 0.299 / 0.302 / 0.093 / 0.253.
- *last-visit-only*: abstention 0.883 / 0.895 / 0.994 / 0.926; over-answer 0.117 / 0.105 / 0.006 / 0.074.

Two findings. First, *full-context* beats blind on every backbone—any patient context lifts abstention. Second, abstention is non-monotonic: *last-visit-only* dominates *full-context* on every backbone by 5–30 pp (Haiku 0.994 vs. 0.907; Sonnet 0.926 vs. 0.747; DeepSeek 0.883 vs. 0.701; GPT-4o-mini 0.895 vs. 0.698). Adding dialogue beyond the most-recent encounter actively degrades abstention. Figure 4 shows the over-answer fraction climbing from *last-visit-only* through *full-context* to *structured-timeline*. GPT-4o-mini's anomalous blind T9 of 0.685 (vs. its pooled-blind 0.332) reflects a default abstention prior under empty context—a backbone-specific blind hedging. Per the trivial-baseline

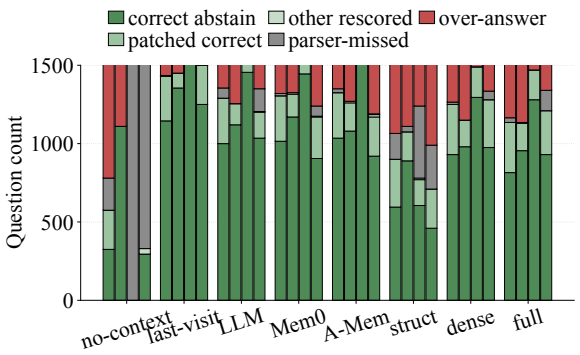

Figure 4: T9 per-cell decomposition into five outcomes: correct abstain (refused as expected), patched correct / other rescored (post-hoc parser recovered abstention via synonyms), parser-missed (verbose abstention not captured), over-answer (confident hallucination); 1500 q/cell, bars are DeepSeek/GPT-4o-mini/Haiku/Sonnet left to right within each strategy cluster. Abstention rate = correct + patched + parser-missed. Over-answer grows from *last-visit-only* to *full-context* and *structured-timeline*: abstention is non-monotonic in context length.

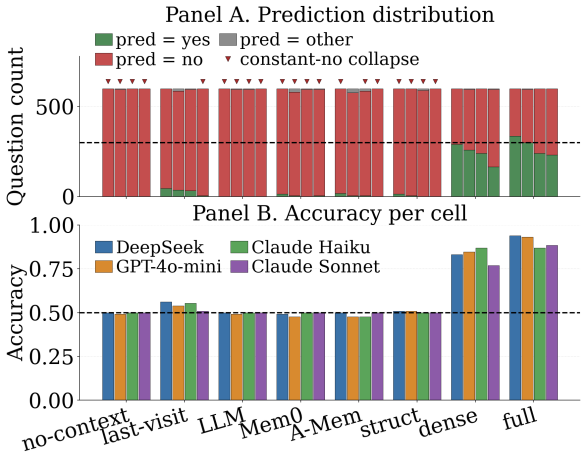

Figure 5: T3 constant-No collapse. (A) Per-cell yes/no/other prediction distribution (600 q/cell); dashed line at 300 = balanced gold; red triangles flag cells with pred = no $\geq 575/600$. (B) Per-cell accuracy; dashed 0.5 reference = constant-No floor. The four compressed strategies collapse to flat 0.50 bars across all backbones (16 cells)— when the sentence-local signal is lost, the model defaults to "No".

caveat (§3), T9 is a relative ordering, not an absolute measure.

5.5 SP4: Controlled preservation probe: compressed representations collapse to the constant-No floor on attribution detection

T3 is the cleanest demonstration of preprocessing-induced signal loss and the strongest finding in this paper. Each "yes" instance has a single attribution sentence injected between a same-encounter finding and its problem; "no" instances are unmodified. The injection is sentence-local: *full-context*, *last-visit-only*, and *dense-retrieval* preserve it; *Mem0*, *A-Mem*, *llm-summary*, and *structured-timeline* (built upstream of injection) cannot. The results align with the design:

- *full-context*: DeepSeek 0.938, GPT-4o-mini 0.931, Haiku 0.869, Sonnet 0.885.
- *dense-retrieval*: DeepSeek 0.831, GPT-4o-mini 0.846, Haiku 0.869, Sonnet 0.769.
- *Mem0* / *A-Mem* / *llm-summary* / *structured-timeline*: 0.477–0.508 across all four backbones.
- *no-context-blind*: 0.492–0.500.

All four compressed strategies × four backbones (16 cells) collapse to "No" on all 600 T3 questions (temperature 0.1, raw responses literally "No"); the balanced 300 / 300 gold yields exactly 0.500 accuracy—the constant-No floor. This is not random guessing but a deterministic failure mode signaling complete loss of the injected signal. Figure 5 visualizes it: Panel A stacks yes / no / other per cell with red triangles flagging constant-No cells; Panel B plots the flat 0.50 band across the 16 cells.

The full-to-compressed gap is +43 to +45 pp—larger than any blind-to-full lift on any other task—and it is not a function of backbone capacity: even GPT-4o-mini, the weakest backbone overall, reaches 0.931 on T3 with full context. The controlled-injection design makes T3 accuracy a pure function of whether the strategy retained the original visit text. The paradigm generalizes: any property encodable as a sentence-level signal can be tested for representation preservation by post-hoc injection.

Caveat. T3 is not an apples-to-apples cross-strategy ranking, since four strategies build representations upstream of injection. It is best read as a preservation probe: Figure 5B reports accuracy for the three downstream strategies and the constant-No floor for the four upstream ones. The clinical motivation is encounter-level evidence–problem linkage, but the actual probe is signal preservation rather than unsupervised linkage reasoning (see §Discussion for re-injection plans).

5.6 Retrieval, cost, and Pareto efficiency

Dense retrieval matches *full-context* accuracy at a fraction of the cost. BGE-M3 hit rates over 4,771 retrieval-eligible questions: hit@1 = 0.595, hit@3 = 0.852, hit@5 = 0.926, hit@10 = 0.979 (Appendix C). Figure 6 places every cell in (cost, accuracy) space and traces the 11-cell Pareto fron-

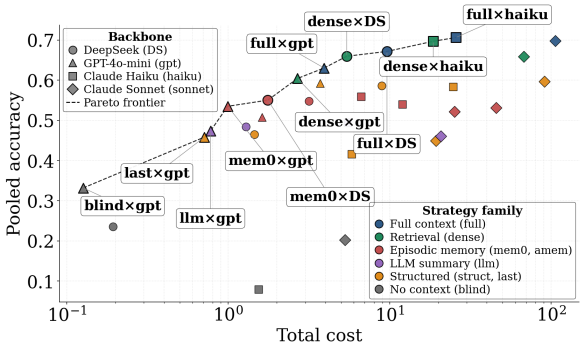

Figure 6: Cost vs. pooled accuracy across 32 cells. Cost: USD per 6,271-question cell (log scale). Color = strategy family; shape = backbone. Dashed curve: Pareto frontier (11/32). Haiku dominates Sonnet on *full-context*; *structured-timeline* is off-frontier on cost despite competitive accuracy.

tier. Two consequences. First, *full-context* × Sonnet ($106.21, 0.698) is dominated by *full-context* × Haiku ($25.76, 0.706), inverting the "biggest backbone wins" heuristic. Second, *structured-timeline* never reaches the frontier despite competitive accuracy (0.58–0.60)—the structured tokens fail to compress enough to offset per-call cost.

5.7 Illustrative failure modes

Aggregate accuracy collapses categorically distinct errors into a single number. Four case studies in Appendix G (Figure 25) make the distinction concrete: T3 constant-No collapse (Case 1), T9 over-answering (Case 2), Haiku-specific T2 trend failure (Case 3), and T6 cross-patient confusion (Case 4). The underlying errors differ in kind—conflating them obscures the diagnostic signal that each representation strategy fails in a characteristic way.

6 Discussion

What compact representations get right. Dense retrieval lands within 1–4 pp of *full-context* pooled and within 0–9 pp on every task except T4 (small-n); when answers are chunk-localizable, BGE-M3 surfaces them reliably. Structured timelines beat *full-context* on numeric retrieval (T1 0.886–0.923) by reducing noise around the cell. *Mem0* and *A-Mem* hold their own on T4 and T9 where compression incidentally aligns with gold density.

What compact representations get wrong. (i) Aggregation (T2/T6b/T6c, exploratory T8): preprocessing discards per-visit value pairs needed for trends/deltas. (ii) Sentence-signal preservation (T3): single-sentence signals are lost in any strategy that builds derived artifacts pre-evaluation—the SP4 constant-No floor. (iii) Contradiction (T5): trivial always-contradicts scores 91.9%; no cell beats it. (iv) Non-monotonic abstention (T9): *full-context* helps over blind but degrades vs. *last-visit-only* on every backbone, and trivial always-abstain scores 100% by construction.

Cost flips "biggest backbone wins". On the Pareto frontier (Figure 6), Haiku not Sonnet sits on it under *full-context*: cheaper ($25.76 vs. $106.21) and slightly more accurate (0.706 vs. 0.698). Memory-equipped clinical agents should be benchmarked on the cost–quality frontier.

Memory size $\neq$ accuracy. Total *Mem0* character budget is uncorrelated with patient-level T1 *Mem0* accuracy (Pearson $r = -0.129$; Appendix 15). Architectural choice (atomic facts vs. visit-level notes vs. prose) matters more than budget; more facts can introduce retrieval noise.

Future work. Multi-EHR replication (eICU-CRD, MIMIC-CXR), longer time horizons, dynamic memory updates, and backbone-matched *Mem0*/*A-Mem* prep. Future releases will rebalance T5/T8 to remove trivial-class ceilings and inject T3 attribution signals before derived-artifact construction, putting every representation on equal footing for preservation.

7 Conclusion

On ClinTraceBench (9 tasks, 385 dialogues, 8 ClinTraceBench provides the first source-verifiable benchmark for longitudinal clinical reasoning over EHR-derived dialogues, evaluating eight history representation strategies against four frontier backbones on 6,271 questions. Three high-level findings emerge. First, full context remains the accuracy upper bound, but the gap to compressed representations is driven not by retrieval failure but by a preprocessing tax that destroys sentence-local signals—a deterministic failure mode that no scaling of backbone capacity recovers. Second, abstention discipline does not improve monotonically with context length, complicating the case for feeding longer charts. Third, the cost–quality Pareto frontier inverts the "biggest backbone wins" heuristic, with smaller backbones dominating under full-context. We release the benchmark, all model predictions, the scoring harness, and the controlled-injection generator to support reproducible evaluation of memory architectures in clinical settings.

Ethics statement

Data access and de-identification. ClinTraceBench is derived from MIMIC-IV (Johnson et al., 2023), a HIPAA-compliant, de-identified electronic health record dataset released under the PhysioNet Credentialed Health Data License. All authors completed the required CITI training and signed the PhysioNet Data Use Agreement before accessing the data. No re-identification was attempted; the synthetic dialogues are generated from structured fields only (dates, lab values, ICD codes, drugs), and contain no protected health information.

Synthetic dialogue scope. The 385 verified dialogues are LLM-generated doctor–patient conversations synthesized from de-identified structured records. They are not real clinical communications, are not clinically validated, and must not be used for clinical decision support, model training intended for deployment, or any patient-facing application.

Release. We release the benchmark question set, gold answers, all 200,672 model predictions, the scoring harness, and the controlled-injection generator at *[URL placeholder]*. Raw dialogue text inherits the redistribution restrictions of MIMIC-IV; credentialed researchers can regenerate dialogues using the released pipeline. The benchmark is released for research evaluation only.

Clinical safety and intended use. ClinTraceBench evaluates information retrieval and reasoning fidelity over EHR-derived dialogues. It does not measure clinical safety, diagnostic accuracy, treatment appropriateness, or patient outcomes, and must not be used as a proxy for clinical readiness of any deployed system.

8 Limitations

Scope. (i) Single EHR source (MIMIC-IV); generalization to other settings, languages, and pediatric populations is untested. (ii) Single time horizon (385 dialogues); multi-year follow-up and dynamic memory updates are deferred. (iii) T5 is controlled rather than natural-EHR contradiction; T8 measures post-treatment lab change rather than causal efficacy (concurrent-medication annotations were not retained); T9 covers unstated lab and medication facts only. (iv) *full-context* serves as a coverage upper bound, not a deployment target, given its cost and latency profile.

Residual artifacts. Approximately 1,400 empty API responses (0.7% of 200,672; Appendix E) are retained in headline numbers: excluding them would inflate Sonnet / Haiku *full-context* by at most $+0.2$ pp (well below all headline contrasts), bias the comparison toward GPT-4o-mini / Haiku retrieval cells, and discard deployment-relevant signal. The sensitivity appendix confirms that all SP1–SP4 contrasts hold ($\Delta < 0.5$ pp) under either policy. Roughly 13% of Sonnet T8 rows hit the original 120-token cap before we extended it; residual mid-sentence truncation may persist on a small subset. Trivial baselines show T5 / T6 / T8 / T9 are not discriminable from majority-class predictors at $\alpha = 0.05$, so T1 / T3 / T4 / T7 carry the cleanest representation signal and T5 / T6 / T8 / T9 serve as supporting probes. Long-context degradation is uniform across backbones (-13 to -14 pp from shortest to longest quartile), but length and disease severity remain entangled.

Agentic-memory prep confound. A fixed DeepSeek-V3 prep model renders non-DeepSeek *Mem0* / *A-Mem* cells a heterogeneous pipeline; SP2's gap on agentic cells therefore partly conflates answering- and construction-side effects. Backbone-matched prep is left to future work; we expect it to reshuffle the agentic ordering in Table 1 but not to overturn SP4, which probes preservation independent of prep identity.

References

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. LongBench: A bilingual, multitask benchmark for long context understanding. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), pages 3119–3137.

Suhana Bedi, Yixing Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A. Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, Hyo Jung Hong, Mehr Kashyap, Akash R. Chaurasia, Nirav R. Shah, Karandeep Singh, Troy Tazbaz, Arnold Milstein, Michael A. Pfeffer, and Nigam H. Shah. 2025. MedHELM: Holistic evaluation of large language models for medical tasks. arXiv preprint arXiv:2505.23802.

Asma Ben Abacha, Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Meliha Yetisgen, Fei Xia, and Thomas Lin. 2024. MEDEC: A benchmark for medical error detection and correction in clinical notes. arXiv preprint arXiv:2412.19260.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu

Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. arXiv preprint arXiv:2402.03216, 4(5).

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. Mem0: Building production-ready AI agents with scalable long-term memory. arXiv preprint arXiv:2504.19413.

Hejie Cui, Alyssa Unell, Bowen Chen, Jason Alan Fries, Emily Alsentzer, Sanmi Koyejo, and Nigam H Shah. 2025. Timer: Temporal instruction modeling and evaluation for longitudinal clinical records. npj Digital Medicine, 8(1):577.

Scott L. Fleming, Alejandro Lozano, William J. Haberkorn, Jenelle A. Jindal, Eduardo P. Reis, Rahul Thapa, Louis Blankemeier, Julian Z. Genkins, Ethan Steinberg, Ashwin Nayak, Birju S. Patel, Chia-Chun Chiang, Alison Callahan, Zepeng Huo, Sergios Gatidis, Scott J. Adams, Oluseyi Fayanju, Shreya J. Shah, Thomas Savage, Ethan Goh, Akshay S. Chaudhari, Nima Aghaeepour, Christopher Sharp, Michael A. Pfeffer, Percy Liang, Jonathan H. Chen, Keith E. Morse, Emma P. Brunskill, Jason A. Fries, and Nigam H. Shah. 2024. MedAlign: A clinician-generated dataset for instruction following with electronic medical records. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 22021–22030.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, Samuel Tesch, Ryan Laffin, Matthew M. Churpek, and Majid Afshar. 2023. DR.BENCH: Diagnostic reasoning benchmark for clinical natural language processing. Journal of Biomedical Informatics, 138:104286.

Yuanzhe Hu, Yu Wang, and Julian McAuley. 2025. Evaluating memory in LLM agents via incremental multi-turn interactions. arXiv preprint arXiv:2507.05257.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. 2023. Mimic-iv, a freely accessible electronic health record dataset. Scientific data, 10(1):1.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221.

Greg Kamradt. 2023. Needle in a haystack – pressure testing LLMs. https://github.com/gkamradt/LLMTest_NeedleInAHaystack. GitHub repository.

Sunjun Kweon, Jiyoun Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwanghyun Kim, Jeewon Yang, Seunghyun Won, and Edward Choi. 2024. Ehrnoteqa: An llm benchmark for real-world clinical practice using discharge summaries. Advances in Neural Information Processing Systems, 37:124575–124611.

Yash Kumar Lal, Claire Cardie, Veselin Stoyanov, and Ana Marasović. 2024. Clinical temporal reasoning evaluation of large language models. arXiv preprint arXiv:2403.01345. Citation details approximate; please verify on arXiv before final submission.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems (NeurIPS), volume 33, pages 9459–9474.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics, 12:157–173.

Mingyu Derek Ma, Chenchen Ye, Yu Yan, Xiaoxuan Wang, Peipei Ping, Timothy S Chang, and Wei Wang. 2024. Clibench: A multifaceted and multigranular evaluation of large language models for clinical decision making. arXiv preprint arXiv:2406.09923.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. 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), pages 13851–13870.

Charles Packer, Vivian Fang, Shishir_G Patil, Kevin Lin, Sarah Wooders, and Joseph_E Gonzalez. 2023. Memgpt: towards llms as operating systems.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. Nature, 620(7972):172–180.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. Journal of the American Medical Informatics Association (JAMIA), 20(5):806–813.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2025. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In International Conference on Learning Representations (ICLR).

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025. A-MEM: Agentic memory for LLM agents. In Advances in Neural Information Processing Systems (NeurIPS).

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2026. A-mem: Agentic memory for llm agents. Advances in Neural Information Processing Systems, 38:17577–17604.

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2025. A survey on the memory mechanism of large language model based agents. ACM Transactions on Information Systems (TOIS). ArXiv:2404.13501.

A Validation gates

Each (anchor, gold) pair passes through five deterministic levels (L0–L4) plus a final human spot-check (L5) before entering the 6,271-question evaluation set:

- **L0 — Schema.** JSONL fields and types match the task-specific schema; threshold 0 failure.
- **L1 — Anchor source.** Anchors for T1–T8 must lie in `verified_covered_events`; T9 anchors must lie in `missed_events`; threshold $< 1\%$.
- **L2 — Gold recompute.** The gold is independently regenerated from the spec and the raw structured fields, and must equal the candidate gold; threshold $< 0.5\%$.
- **L3 — No gold leakage.** The question text must not contain the gold answer (critical for T2 trend labels and T9 abstention cues); threshold 0 failure.
- **L4 — Distribution coverage.** Per-task disease, complexity-tertile, patient, and event-type distributions are checked against the construction targets; reported as a soft warning rather than a hard fail.
- **L5 — Human spot-check.** A stratified sample of 370 questions (30/task low-risk $\times$ 4 + 50/task high-risk $\times$ 5) was hand-audited by the authors; 366/370 = **98.92%** agreement with the automated gold. 7/9 tasks scored 0% FAIL; T7 = 2% (distractor substring leak), T9 = 6% (Secondary-pool textual templating). Both below the 10% fix threshold.

L0–L4 are deterministic and run on all generated questions; L5 is a human verification pass on a stratified subsample.

B Per-task subtype diagnostics

The headline tables pool over every within-task subtype for compactness, but each of our nine evaluation tasks has at least one principled axis along which difficulty varies by construction. This appendix reopens those axes one task at a time, so the reader can see which question variants drive the aggregate task scores in the main paper. Three recurring patterns are worth flagging up front. First, tasks whose subtypes are categorical anchors (T3 attribution class, T7 medication vs procedure, T9 silent-evidence kind) split cleanly into separate distributions. Second, tasks with an explicit controlled-vs-natural design (T5) collapse onto that family axis and the headline finding lives there. Third, tasks with a magnitude $\times$ direction lattice (T8) need both axes to read off difficulty; small-magnitude bins are categorically harder than larger ones across every backbone.

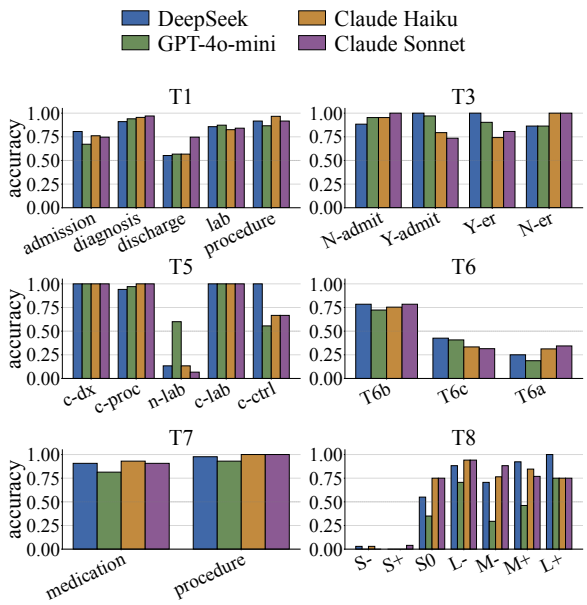

Figure 7: Per-task subtype accuracy under full-context across four backbones. T5 short codes: c-dx / c-proc / c-lab / c-ctrl = controlled_dx_presence / _procedure / _lab_value / _control; n-lab = natural_c2_lab_non_glucose. T8 short codes: L/M/S = large/medium/small magnitude; −/+/0 = negative/positive/flat direction.

B.1 Subtype overview

Figure 7 is the six-panel overview: T1, T3, T5, T6, T7, T8 under full-context across all four backbones. T2 / T4 / T9 are deferred to their own dedicated figures below because their subtype axes do not match the generic layout. Per-task subtype n counts are listed inside the figure.

B.2 T3 four-class attribution detail

T3's design crosses injection presence (**Y / N**) with encounter category (**admit / er**) to produce four classes. The DeepSeek panel in Figure 8 is the representative split; the other three backbones show the same qualitative pattern. The story is the same in every backbone we tested: injected cells (Y-$*$) collapse to $\sim$0 under compressed memories and are recovered by `dense` and `full-context`; not-injected cells (N-$*$) sit near ceiling for every strategy except the dense / full-context strategies, where additional context occasionally adds noise.

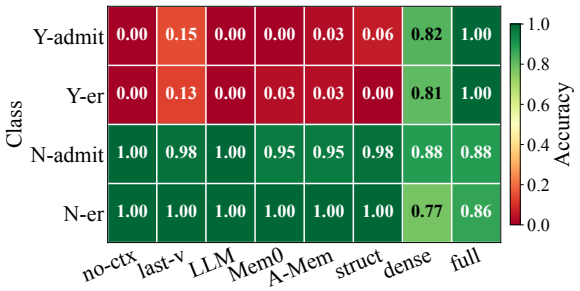

Figure 8: T3 attribution-detection accuracy by 4-class stratification (DeepSeek backbone, representative). n: Y-admit 34, Y-er 31, N-admit 43, N-er 22.

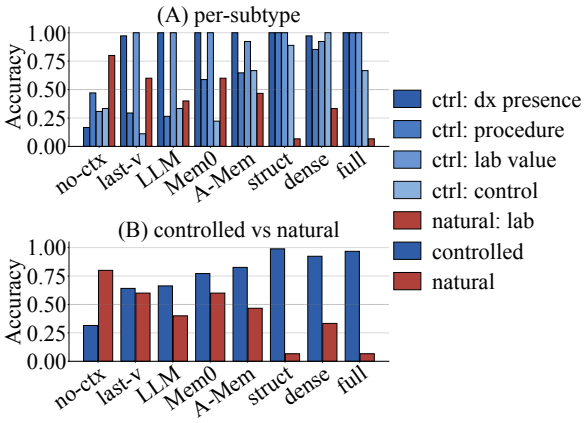

Figure 9: T5 controlled-vs-natural contradiction breakdown (Sonnet backbone exemplar). (A) per-subtype accuracy across the 8 memory baselines; (B) the same data collapsed onto the controlled-vs-natural family axis.

B.3 T5 controlled vs. natural contradiction

T5 is designed as a controlled contradiction probe rather than a natural EHR-contradiction mining task. Figure 9 makes that explicit: the five subtypes collapse cleanly onto a single controlled-vs-natural axis (controlled $n = 92$, natural $n = 15$), and the two families move in opposite directions as memory becomes richer. Controlled subtypes climb from 0.32 at `no-ctx` to 0.97 under `full-context`, while the lone natural subtype (lab non-glucose) goes the other way (0.80 $\rightarrow$ 0.07), because `dense` and `full-context` introduce confounding lab evidence that triggers spurious contradictions.

B.4 T4 retrospective evidence retrieval

T4 scores strict set-F1 on event UIDs and is the smallest task in the suite ($n = 53$); the question pool is biased toward a handful of condition families. Figure 10 reports the pool composition, per-backbone full-context set-F1, and the full 8×4 cell heatmap. The small-n profile is intentional and is the reason T4 is treated as a limited rather than headline finding throughout the paper.

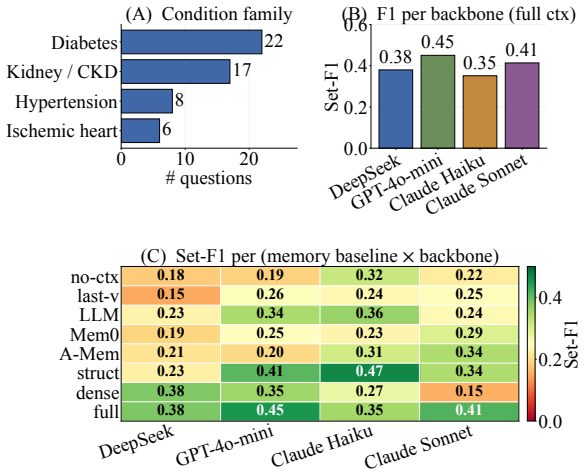

Figure 10: T4 retrospective evidence retrieval. (A) condition-family distribution of the $n = 53$ questions; (B) per-backbone set-F1 under full-context; (C) set-F1 heatmap across 8 memory baselines $\times$ 4 backbones.

B.5 T8 observed treatment-response bins

T8 is observed post-treatment lab change tracking, not causal treatment-response estimation; concurrent-medication annotations were deliberately not retained in the final evaluation parquet, so the seven magnitude $\times$ direction bins are a difficulty proxy rather than a clinical-effect estimate. The headline pattern is that large/medium bins climb from 0.29–0.65 at `no-ctx` to 0.69–0.94 at `full-context`, but the small bins are categorically harder: **S**$-$ and **S**$+$ sit at $\sim$0 across every baseline including `full-context`, while **S**0 is rescued only by `structured timeline` (1.00) and `full-context` (0.75).

B.6 T9 silent-evidence kind: lab vs. other

The active T9 pool is structurally degenerate along the primary-or-secondary axis: every question has `anchor_count` $= 1$ and `primary_or_secondary` = "primary". The cleanest available stratification is therefore the silent-evidence kind (lab vs. other), and Figure 12 reports accuracy, abstention rate, and over-answer rate on that split. The 246 lab-anchored vs. 78 other-anchored split inside the active pool preserves the same $\approx$ 76%/24% ratio as the underlying T9 generation pool (1,147 vs. 353).

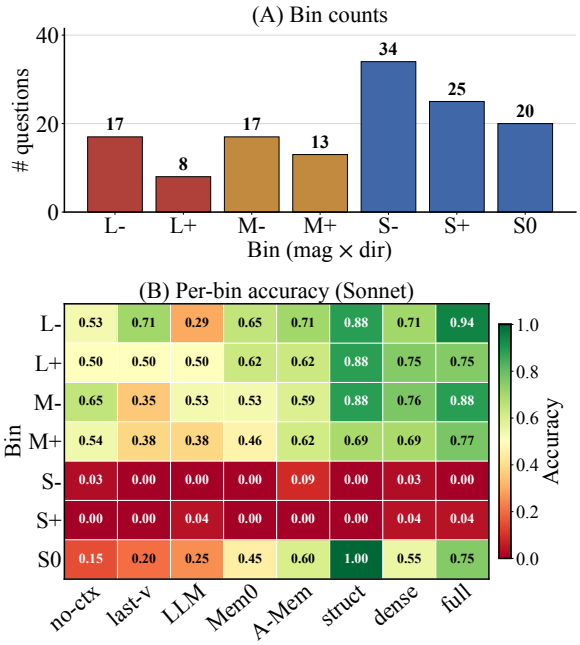

Figure 11: T8 per-bin breakdown (Sonnet exemplar). (A) question counts in the seven (magnitude × direction) bins ($n = 134$ total); (B) per-bin accuracy across the 8 memory baselines.

C Cohort, retrieval, and compression footprints

This appendix characterises the static building blocks of the benchmark: the patient cohort it draws from, the auxiliary retrieval index it lets the `dense` baseline rely on, and the compressed-memory structures the `Mem0` / `A-Mem` / `LLM summary` baselines build upstream of answer time. The point is to make clear that the headline accuracy differences come from representation choice, not from asymmetric access to the underlying clinical evidence.

C.1 Patient cohort

The 6,271-question evaluation set is drawn from $n = 385$ verified dialogues. Figure 13 reports the three distributional cuts that matter: visits per patient (median 10, mean 16.0, max 93), dialogue length in thousands of characters (median 13.7 k, mean 15.0 k, max 30.7 k), and stratified-eval questions per patient (median 4, mean 4.3, max 12). The vast majority of the 385 dialogues contribute at least one question to the eval set.

C.2 BGE-M3 dense-retrieval hit rate

The `dense` baseline relies on BGE-M3 to embed visits and rank them against the query. Figure 14 reports top-k visit-hit rate per task on the 1,177 retrieval-eligible questions (T9 silent-evidence is excluded by construction). The overall headline is hit@1 $= 0.595$, hit@3 $= 0.852$, hit@5 $= 0.926$, hit@10 $= 0.979$, which is sufficient quality that the retrieval step is not the bottleneck on any of the tasks where `dense` loses to `full-context`.

C.3 Compression footprints

The three compressed-memory strategies produce structures of very different sizes. Figure 15 reports the per-patient footprint of each, plus a size-vs-accuracy scatter that disposes of the hypothesis that headline accuracy differences are driven by raw representation size. Pearson $r(\texttt{Mem0}\ \text{size}, \texttt{Mem0}\ \text{T1 accuracy}) = -0.129$ on $n = 210$ patients: bigger is not better, and the compressed representations are bottlenecked by what they encode, not by how much.

D Resource consumption

Tables and figures in this appendix attach a dollar / token / cache cost to each of the 32 (memory baseline × backbone) cells. The takeaway is the same one the main paper draws: most of the representation-driven accuracy gains do not require the most expensive cell, and the cost ranges across the grid are wide enough that efficiency–accuracy trade-offs are a real decision rather than a marginal one.

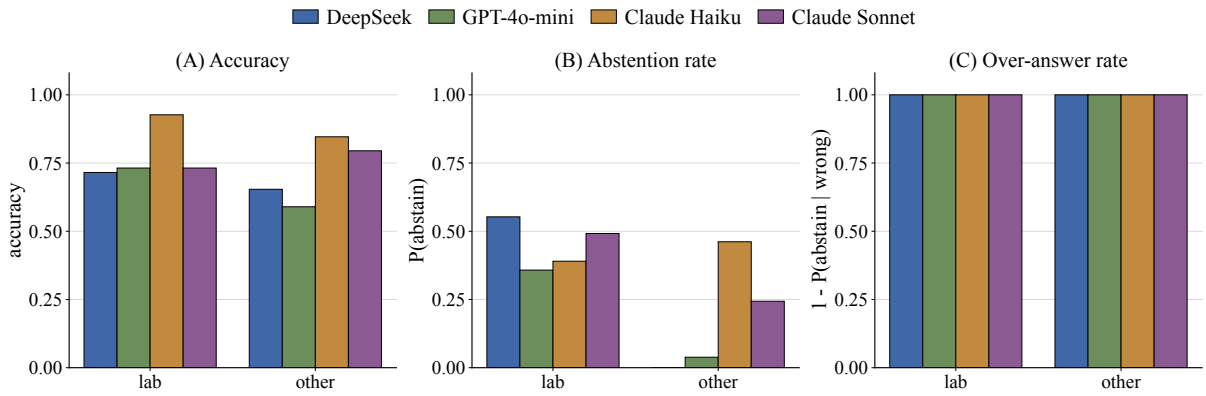

Figure 12: T9 silent-evidence abstention stratified by anchor kind (lab vs. other) under full-context. (A) accuracy; (B) abstention rate; (C) over-answer rate when wrong. $n = 246$ lab vs. $n = 78$ other.

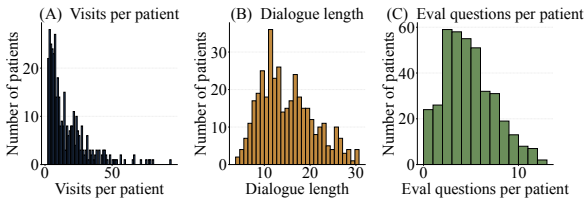

Figure 13: Cohort statistics for the $n = 385$ verified dialogues. (A) visits per patient, (B) dialogue length, (C) eval questions per patient.

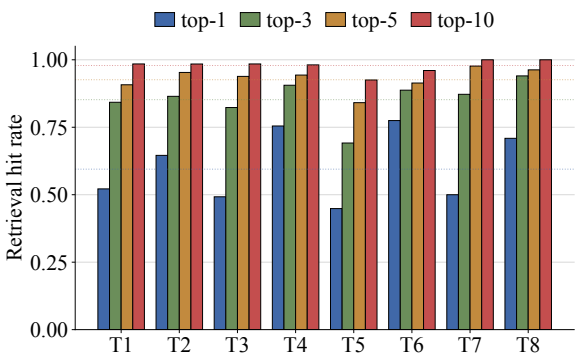

Figure 14: BGE-M3 dense-retrieval hit rate by task on 1,177 retrieval-eligible questions. Per-task eligible n: T1 324, T2 192, T3 130, T4 53, T5 107, T6 151, T7 86, T8 134. Dotted lines mark the overall hit rate at each k.

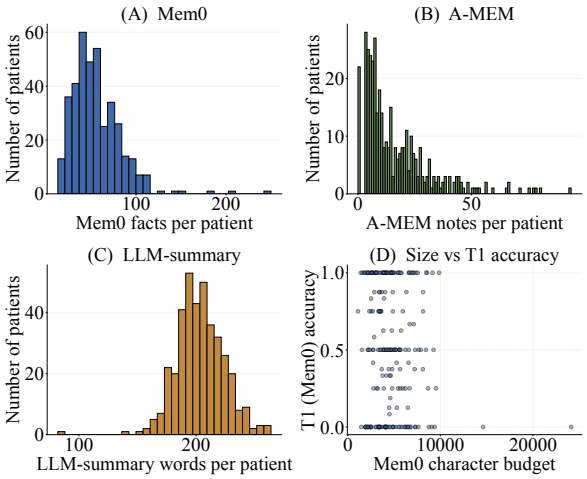

Figure 15: Per-patient compression footprints. (A) `Mem0` facts per patient, (B) `A-Mem` notes per patient, (C) `LLM-summary` words per patient, (D) total `Mem0` size vs T1 accuracy ($n = 210$, Pearson $r = -0.129$).

Table 3: **Per-cell resource consumption projected to the full 6,271-question evaluation set.** Each row is one (strategy, backbone) cell. Cost (USD) and tokens are linearly projected from the 1,501-question stratified pilot ($\times 4.178$); cache hit % is pilot-observed and is rate-stable across scaling. Per-question cost ($\$/q$) is cost / 6,271. Total projected spend: $527.87 across all 32 cells.

Strategy	Backbone	Cost (USD)	$/q	Input (M)	Output (K)	Cache hit %
no-context blind	DeepSeek-V3	$0.19	$0.0000	0.5	52.3	2.2%
	GPT-4o-mini	$0.13	$0.0000	0.5	74.3	0.0%
	Haiku 4.5	$1.54	$0.0002	0.6	272.8	0.0%
	Sonnet 4.6	$5.27	$0.0008	0.6	238.1	0.0%
last-visit-only	DeepSeek-V3	$1.45	$0.0002	5.3	40.4	1.9%
	GPT-4o-mini	$0.71	$0.0001	5.5	81.9	21.9%
	Haiku 4.5	$5.80	$0.0009	6.4	208.1	3.0%
	Sonnet 4.6	$19.34	$0.0031	6.4	163.1	13.9%
full-context	DeepSeek-V3	$9.62	$0.0015	36.8	51.3	4.1%
	GPT-4o-mini	$3.92	$0.0006	37.7	78.8	35.0%
	Haiku 4.5	$25.76	$0.0041	44.7	137.9	32.8%
	Sonnet 4.6	$106.20	$0.0169	44.9	141.4	25.2%
dense-retrieval	DeepSeek-V3	$5.42	$0.0009	20.1	52.2	1.4%
	GPT-4o-mini	$2.68	$0.0004	20.2	75.2	14.4%
	Haiku 4.5	$18.61	$0.0030	24.0	131.4	6.5%
	Sonnet 4.6	$67.63	$0.0108	24.2	131.6	10.8%
structured-timeline	DeepSeek-V3	$8.91	$0.0014	34.2	45.2	4.4%
	GPT-4o-mini	$3.70	$0.0006	34.9	69.5	33.4%
	Haiku 4.5	$24.72	$0.0039	41.0	158.1	29.6%
	Sonnet 4.6	$90.83	$0.0145	41.2	133.2	31.3%
LLM-summary	DeepSeek-V3	$1.29	$0.0002	4.7	50.3	2.9%
	GPT-4o-mini	$0.78	$0.0001	4.9	78.9	1.2%
	Haiku 4.5	$5.49	$0.0009	5.9	187.9	0.0%
	Sonnet 4.6	$20.78	$0.0033	6.0	200.0	0.5%
mem0	DeepSeek-V3	$1.76	$0.0003	6.4	45.3	0.4%
	GPT-4o-mini	$0.99	$0.0002	6.4	75.3	0.8%
	Haiku 4.5	$6.65	$0.0011	7.5	158.9	0.0%
	Sonnet 4.6	$25.16	$0.0040	7.7	145.0	0.0%
amem	DeepSeek-V3	$3.17	$0.0005	11.7	48.1	1.7%
	GPT-4o-mini	$1.62	$0.0003	12.0	76.1	14.1%
	Haiku 4.5	$11.99	$0.0019	14.1	168.1	0.0%
	Sonnet 4.6	$45.75	$0.0073	14.5	143.5	0.0%
Total	—	**$527.87**	—	**532**	**3714**	**17.1%**

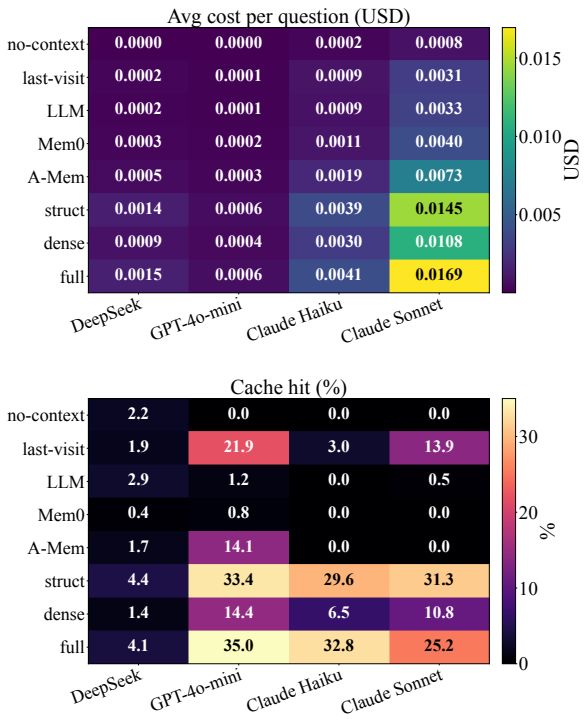

Figure 16: Companion visualisation to Table 3: avg cost per question (top) and cache hit % (bottom) per (strategy, backbone) cell. Sonnet × `full-context` is the most expensive cell ($0.0169/q); DeepSeek × `no-context blind` is the cheapest ($0.00003/q). Cache-hit range: 0%–35%.

E Pipeline-quality diagnostics

This appendix documents the failure modes of the evaluation pipeline itself. Two kinds of issue can in principle confound the headline accuracies: a model returning no response at all, and the strict gold parser failing to recover a prediction. We measure both, and we measure where each one concentrates so the reader can see they are not driving the comparisons in the main paper.

E.1 Empty responses (D1)

After the M1 retry pass, 336 of the 48,032 prediction rows (0.7%) still have an empty `raw_response`. They concentrate on GPT-4o-mini ($n = 160$) and Claude Haiku ($n = 139$) under retrieval-style contexts (`Mem0`, `A-Mem`, `dense`); DeepSeek's column is uniformly zero. These rows are documented but not excluded from the headline accuracy numbers.

E.2 Parse failures (D2)

A response can be present but un-parseable by the strict gold-format matcher. 3,139 / 48,032 rows (∼6.5%) fall into this bucket, and

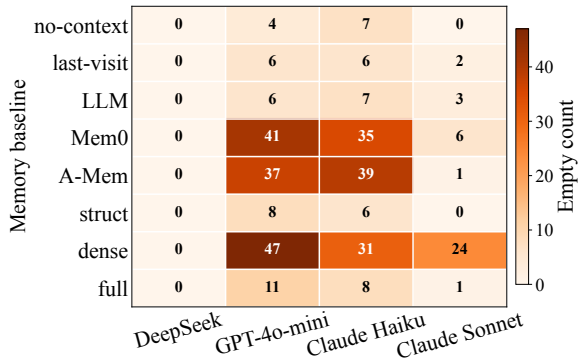

Figure 17: D1: empty API responses per cell after the M1 retry pass (336 / 48,032 rows; per-backbone totals: DeepSeek 0, GPT-4o-mini 160, Haiku 139, Sonnet 37).

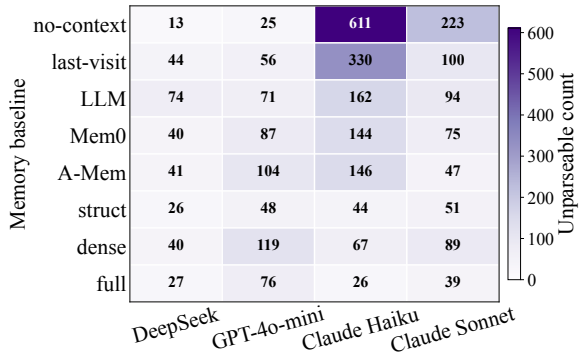

Figure 18: D2: parse-failure rate (`pred is None`) per cell across 8 memory baselines × 4 backbones (1,501 questions per cell).

they are dominated by Claude refusal-as-prose under `no-context blind` – the Haiku × `no-context` cell alone contains 611 parse failures. The pattern is concentrated on a small number of cells and does not affect the comparative ordering between memory strategies under richer contexts.

E.3 Output length and truncation

The model is given a per-task `max_tokens` budget. Two diagnostics follow: the full output-length distribution per task overlaid by backbone (Figure 19), and the rate at which a response is truncated at ceiling minus one token (Figure 20). T4 is the only task where the 120-token ceiling binds in a meaningful way (86–88% truncation on Haiku / Sonnet under the worst memory baseline); T1 and T8 are the next largest binders. T3 / T5 / T6 / T7 are essentially un-truncated across the board.

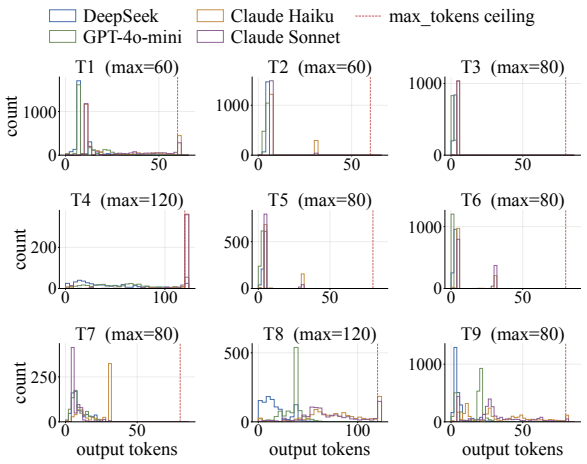

Figure 19: D3: output-length histograms per task overlaid by backbone, with the per-task `max_tokens` ceiling marked.

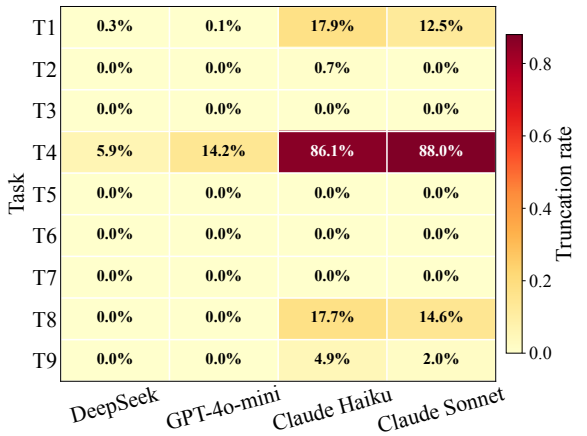

Figure 20: D4: truncation rate per (task, backbone), mean across the 8 memory baselines. Per-task ceilings: T1 60, T2 60, T4 120, T8 120, T9 80; others 80.

F Robustness and stratification

The headline numbers in the main paper are point estimates averaged over patients, diseases, and within-task difficulty tertiles. This appendix reports the four sensitivity checks that we think matter most: per-patient skew, per-disease drift, per-tertile calibration of the stratifier, and bootstrap-CI width per cell.

F.1 Per-patient accuracy distribution (D5)

Figure 21 renders the per-cell distribution of per-patient mean accuracy as a violin. The headline observation is that `full-context` violins are uniformly narrower than the compressed-representation cells on every backbone – richer memory produces less between-patient skew, not just a higher mean.

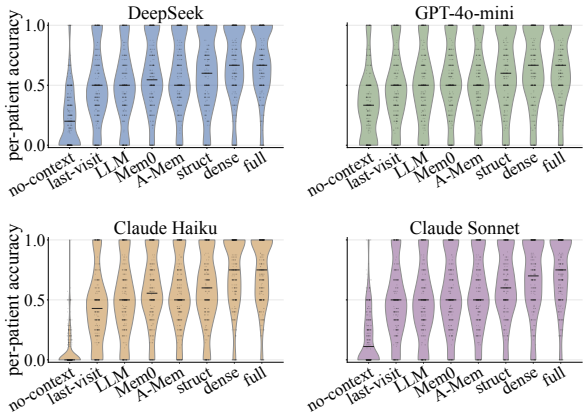

Figure 21: D5: per-patient accuracy distribution as a violin per cell; medians overlaid in black, patient-level scatter in grey.

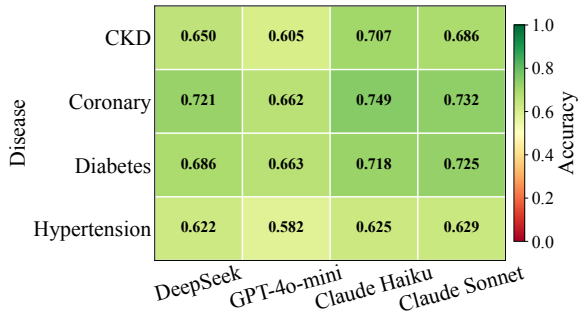

Figure 22: D6: per-disease `full-context` accuracy by backbone, pooled across T1–T9. Per-disease n per backbone: CKD 488, Coronary 358, Diabetes 404, Hypertension 251.

F.2 Per-disease drift (D6)

The benchmark spans four condition families: CKD, Coronary, Diabetes, and Hypertension. Figure 22 reports full-context accuracy pooled across all nine tasks, by backbone $\times$ disease. The hypertension row is consistently the weakest (mean 0.615), but this is a task-mix bias – the hypertension question pool over-indexes the longitudinal / comparison families – rather than evidence of a disease-specific knowledge gap.

F.3 Per-tertile calibration of the stratifier (D7)

Each question carries a difficulty tertile from the generation-time stratifier. Figure 23 checks whether the tertile actually predicts difficulty under `full-context`. T1 and T9 are monotonically calibrated, while T8's "high" tertile is the easiest of the three – T8 tertile anti-alignment, which we flag in the main paper.

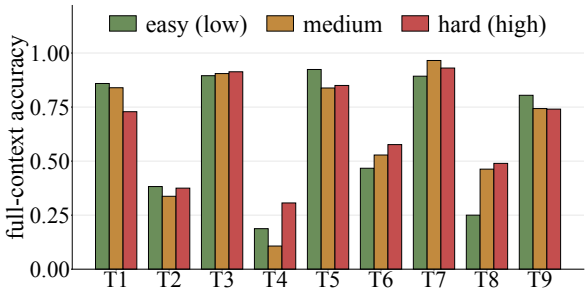

Figure 23: D7: per-tertile `full-context` accuracy per task, pooled across backbones. Per-task tertile n (low/medium/high): T1 440/436/420, T2 68/252/448, T3 124/200/196, T4 32/56/124, T5 92/136/200, T6 212/248/144, T7 84/116/144, T8 144/296/96, T9 440/436/420.

F.4 Bootstrap CI widths per cell (D8)

Figure 24 reports the bootstrap 95% CI width for each of the $9 \times 8 \times 4 = 288$ cells (100 resamples per cell, seed 20260519). The headline is that CI width is dominated by per-cell sample size, not by memory strategy or backbone: T4 (smallest n) and T8 are the widest by construction; T1 / T7 / T9 (the highest-resampled tasks) are the tightest. Conclusion: the headline rankings in the main paper are robust to within-cell sampling noise.

G Case studies

The main paper carries four illustrative failure-mode case studies; we reprint them here so the appendix is self-contained.

H Full per-task strategy-by-backbone matrix

For completeness, the full $9 \times 8 \times 4 = 288$-cell accuracy matrix from which all per-task figures in the main paper and in this appendix are derived is reproduced below.

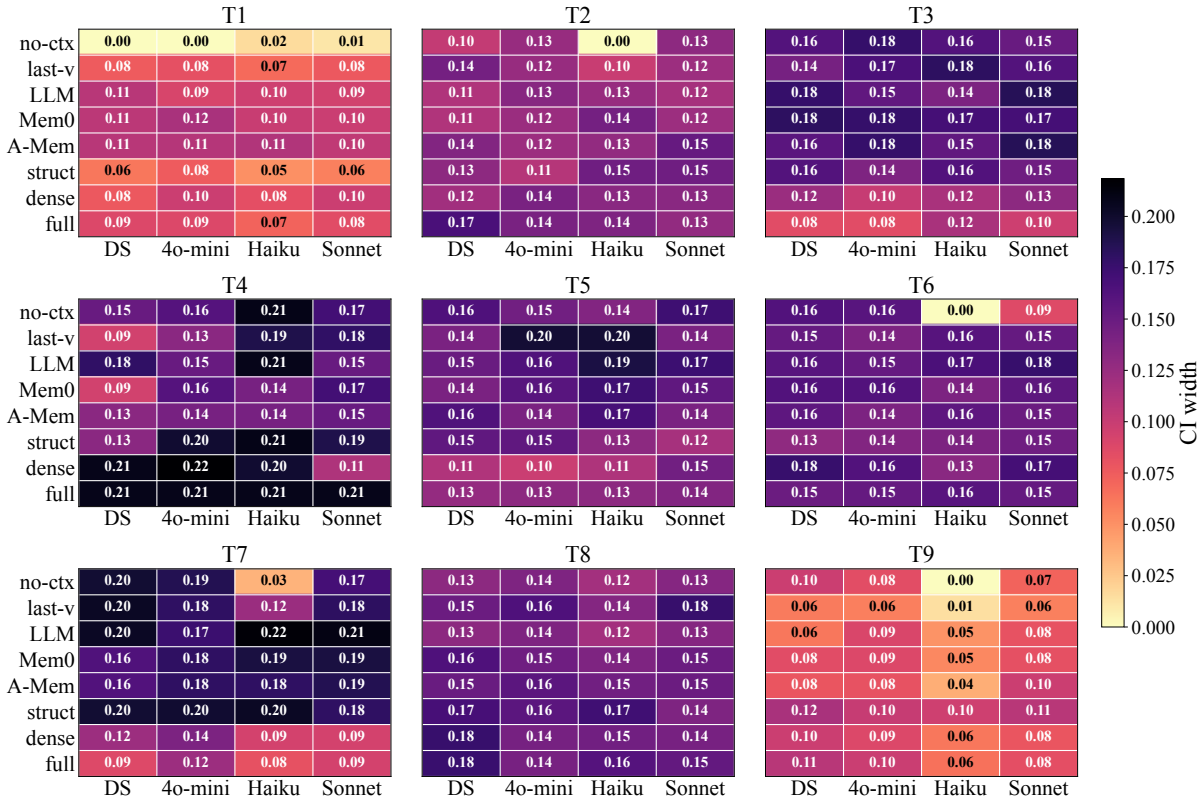

Figure 24: D8: bootstrap 95% CI width per (task, baseline, backbone) cell. Shared `magma_r` colour scale across all nine sub-panels. Per-task questions/cell and mean CI width: T1 ($n = 324$, 0.08), T2 (192, 0.12), T3 (130, 0.15), T4 (53, 0.17), T5 (107, 0.15), T6 (151, 0.15), T7 (86, 0.16), T8 (134, 0.15), T9 (324, 0.08).

Case 1 – T3: Controlled Evidence–Problem Linkage

Q: Did the physician explicitly attribute the Creatinine 1.7 mg/dL (2183-01-13) to a diagnosis of...

(full-context, DeepSeek): yes [correct]

(mem0, DeepSeek): no [incorrect]

Lesson: Compressed representations lose the encounter-local linkage sentence and collapse to constant-No.

Case 2 – T9: Abstention over Unstated Facts

Q: Based on the dialogue, what was the patient's Potassium on 2138-04-03?

(last-visit-only, Sonnet): insufficient [correct]

(full-context, Sonnet): the patient's Pota... [incorrect]

Lesson: More context can hurt abstention; full-context over-answers.

Case 3 – T2: Multi-Visit Trend Classification

Q: How did the patient's Potassium change overall from 2126-02-06 to 2126-12-06? Answer `<direction>;<magnitude>`.

(full-context, Sonnet): decreased;weak [correct]

(full-context, Haiku): stable;none [incorrect]

Lesson: Even with full chart access, multi-visit trend aggregation remains backbone-sensitive.

Case 4 – T6: Cross-Patient Comparison

Q: Comparing patient A and patient B, which patient has more diagnoses documented in the dialogue? Answer A, B, or similar.

(full-context, Sonnet): B [correct]

(mem0, Sonnet): A [incorrect]

Lesson: Agentic memory representations can confuse facts across patients when retrieval is keyed on similar attributes.

Figure 25: Four illustrative failure modes drawn from the master results table: representational loss (Case 1, T3), uncritical answering (Case 2, T9), reasoning-capacity ceiling (Case 3, T2 on Haiku), cross-document aggregation breakdown (Case 4, T6). Green = correct prediction; red = incorrect.

Table 4: Complete per-task accuracy across all 32 (history representation strategy, backbone) cells, on the 6,271-question stratified evaluation set (200,672 evaluations total). `n` is the per-task sample size; per-task sizes are T1=1500, T2=889, T3=600, T4=53, T5=495, T6=700, T7=400, T8=134, T9=1500. The rightmost column marks rows from the stratified evaluation set (every row is).

Task	Strategy	Backbone	n	Accuracy	Subset
T1	no-context (blind)	DeepSeek	1500	0.000	✓
T1	no-context (blind)	GPT-4o-mini	1500	0.000	✓
T1	no-context (blind)	Claude Haiku	1500	0.006	✓
T1	no-context (blind)	Claude Sonnet	1500	0.003	✓
T1	last-visit only	DeepSeek	1500	0.160	✓
T1	last-visit only	GPT-4o-mini	1500	0.148	✓
T1	last-visit only	Claude Haiku	1500	0.127	✓
T1	last-visit only	Claude Sonnet	1500	0.120	✓
T1	LLM summary	DeepSeek	1500	0.355	✓
T1	LLM summary	GPT-4o-mini	1500	0.361	✓
T1	LLM summary	Claude Haiku	1500	0.352	✓
T1	LLM summary	Claude Sonnet	1500	0.355	✓
T1	Mem0	DeepSeek	1500	0.515	✓
T1	Mem0	GPT-4o-mini	1500	0.515	✓
T1	Mem0	Claude Haiku	1500	0.509	✓
T1	Mem0	Claude Sonnet	1500	0.488	✓
T1	A-Mem	DeepSeek	1500	0.491	✓
T1	A-Mem	GPT-4o-mini	1500	0.420	✓
T1	A-Mem	Claude Haiku	1500	0.426	✓
T1	A-Mem	Claude Sonnet	1500	0.485	✓
T1	structured timeline	DeepSeek	1500	0.886	✓
T1	structured timeline	GPT-4o-mini	1500	0.886	✓
T1	structured timeline	Claude Haiku	1500	0.920	✓
T1	structured timeline	Claude Sonnet	1500	0.923	✓
T1	dense retrieval	DeepSeek	1500	0.802	✓
T1	dense retrieval	GPT-4o-mini	1500	0.738	✓
T1	dense retrieval	Claude Haiku	1500	0.778	✓
T1	dense retrieval	Claude Sonnet	1500	0.787	✓
T1	full context	DeepSeek	1500	0.806	✓
T1	full context	GPT-4o-mini	1500	0.781	✓
T1	full context	Claude Haiku	1500	0.812	✓
T1	full context	Claude Sonnet	1500	0.843	✓
T2	no-context (blind)	DeepSeek	889	0.172	✓
T2	no-context (blind)	GPT-4o-mini	889	0.333	✓
T2	no-context (blind)	Claude Haiku	889	0.000	✓
T2	no-context (blind)	Claude Sonnet	889	0.302	✓
T2	last-visit only	DeepSeek	889	0.328	✓
T2	last-visit only	GPT-4o-mini	889	0.354	✓
T2	last-visit only	Claude Haiku	889	0.177	✓
T2	last-visit only	Claude Sonnet	889	0.354	✓
T2	LLM summary	DeepSeek	889	0.260	✓
T2	LLM summary	GPT-4o-mini	889	0.333	✓
T2	LLM summary	Claude Haiku	889	0.323	✓
T2	LLM summary	Claude Sonnet	889	0.339	✓

(continued on next page)

(continued from previous page)

Task	Strategy	Backbone	n	Accuracy	Subset
T2	Mem0	DeepSeek	889	0.333	✓
T2	Mem0	GPT-4o-mini	889	0.297	✓
T2	Mem0	Claude Haiku	889	0.349	✓
T2	Mem0	Claude Sonnet	889	0.391	✓
T2	A-Mem	DeepSeek	889	0.333	✓
T2	A-Mem	GPT-4o-mini	889	0.323	✓
T2	A-Mem	Claude Haiku	889	0.365	✓
T2	A-Mem	Claude Sonnet	889	0.380	✓
T2	structured timeline	DeepSeek	889	0.297	✓
T2	structured timeline	GPT-4o-mini	889	0.214	✓
T2	structured timeline	Claude Haiku	889	0.292	✓
T2	structured timeline	Claude Sonnet	889	0.406	✓
T2	dense retrieval	DeepSeek	889	0.354	✓
T2	dense retrieval	GPT-4o-mini	889	0.266	✓
T2	dense retrieval	Claude Haiku	889	0.427	✓
T2	dense retrieval	Claude Sonnet	889	0.432	✓
T2	full context	DeepSeek	889	0.375	✓
T2	full context	GPT-4o-mini	889	0.255	✓
T2	full context	Claude Haiku	889	0.349	✓
T2	full context	Claude Sonnet	889	0.474	✓
T3	no-context (blind)	DeepSeek	600	0.500	✓
T3	no-context (blind)	GPT-4o-mini	600	0.492	✓
T3	no-context (blind)	Claude Haiku	600	0.500	✓
T3	no-context (blind)	Claude Sonnet	600	0.500	✓
T3	last-visit only	DeepSeek	600	0.562	✓
T3	last-visit only	GPT-4o-mini	600	0.538	✓
T3	last-visit only	Claude Haiku	600	0.554	✓
T3	last-visit only	Claude Sonnet	600	0.508	✓
T3	LLM summary	DeepSeek	600	0.500	✓
T3	LLM summary	GPT-4o-mini	600	0.492	✓
T3	LLM summary	Claude Haiku	600	0.500	✓
T3	LLM summary	Claude Sonnet	600	0.500	✓
T3	Mem0	DeepSeek	600	0.492	✓
T3	Mem0	GPT-4o-mini	600	0.477	✓
T3	Mem0	Claude Haiku	600	0.500	✓
T3	Mem0	Claude Sonnet	600	0.500	✓
T3	A-Mem	DeepSeek	600	0.500	✓
T3	A-Mem	GPT-4o-mini	600	0.477	✓
T3	A-Mem	Claude Haiku	600	0.477	✓
T3	A-Mem	Claude Sonnet	600	0.500	✓
T3	structured timeline	DeepSeek	600	0.508	✓
T3	structured timeline	GPT-4o-mini	600	0.508	✓
T3	structured timeline	Claude Haiku	600	0.500	✓
T3	structured timeline	Claude Sonnet	600	0.500	✓
T3	dense retrieval	DeepSeek	600	0.831	✓
T3	dense retrieval	GPT-4o-mini	600	0.846	✓
T3	dense retrieval	Claude Haiku	600	0.869	✓

(continued on next page)

(continued from previous page)

Task	Strategy	Backbone	n	Accuracy	Subset
T3	dense retrieval	Claude Sonnet	600	0.769	✓
T3	full context	DeepSeek	600	0.938	✓
T3	full context	GPT-4o-mini	600	0.931	✓
T3	full context	Claude Haiku	600	0.869	✓
T3	full context	Claude Sonnet	600	0.885	✓
T4	no-context (blind)	DeepSeek	53	0.132	✓
T4	no-context (blind)	GPT-4o-mini	53	0.075	✓
T4	no-context (blind)	Claude Haiku	53	0.189	✓
T4	no-context (blind)	Claude Sonnet	53	0.113	✓
T4	last-visit only	DeepSeek	53	0.057	✓
T4	last-visit only	GPT-4o-mini	53	0.113	✓
T4	last-visit only	Claude Haiku	53	0.170	✓
T4	last-visit only	Claude Sonnet	53	0.113	✓
T4	LLM summary	DeepSeek	53	0.151	✓
T4	LLM summary	GPT-4o-mini	53	0.170	✓
T4	LLM summary	Claude Haiku	53	0.189	✓
T4	LLM summary	Claude Sonnet	53	0.113	✓
T4	Mem0	DeepSeek	53	0.038	✓
T4	Mem0	GPT-4o-mini	53	0.132	✓
T4	Mem0	Claude Haiku	53	0.094	✓
T4	Mem0	Claude Sonnet	53	0.151	✓
T4	A-Mem	DeepSeek	53	0.075	✓
T4	A-Mem	GPT-4o-mini	53	0.113	✓
T4	A-Mem	Claude Haiku	53	0.094	✓
T4	A-Mem	Claude Sonnet	53	0.132	✓
T4	structured timeline	DeepSeek	53	0.057	✓
T4	structured timeline	GPT-4o-mini	53	0.208	✓
T4	structured timeline	Claude Haiku	53	0.245	✓
T4	structured timeline	Claude Sonnet	53	0.170	✓
T4	dense retrieval	DeepSeek	53	0.208	✓
T4	dense retrieval	GPT-4o-mini	53	0.226	✓
T4	dense retrieval	Claude Haiku	53	0.151	✓
T4	dense retrieval	Claude Sonnet	53	0.057	✓
T4	full context	DeepSeek	53	0.226	✓
T4	full context	GPT-4o-mini	53	0.283	✓
T4	full context	Claude Haiku	53	0.189	✓
T4	full context	Claude Sonnet	53	0.245	✓
T5	no-context (blind)	DeepSeek	495	0.252	✓
T5	no-context (blind)	GPT-4o-mini	495	0.393	✓
T5	no-context (blind)	Claude Haiku	495	0.168	✓
T5	no-context (blind)	Claude Sonnet	495	0.383	✓
T5	last-visit only	DeepSeek	495	0.785	✓
T5	last-visit only	GPT-4o-mini	495	0.579	✓
T5	last-visit only	Claude Haiku	495	0.374	✓
T5	last-visit only	Claude Sonnet	495	0.636	✓
T5	LLM summary	DeepSeek	495	0.832	✓
T5	LLM summary	GPT-4o-mini	495	0.673	✓

(continued on next page)

(continued from previous page)

Task	Strategy	Backbone	n	Accuracy	Subset
T5	LLM summary	Claude Haiku	495	0.636	✓
T5	LLM summary	Claude Sonnet	495	0.626	✓
T5	Mem0	DeepSeek	495	0.804	✓
T5	Mem0	GPT-4o-mini	495	0.729	✓
T5	Mem0	Claude Haiku	495	0.720	✓
T5	Mem0	Claude Sonnet	495	0.748	✓
T5	A-Mem	DeepSeek	495	0.776	✓
T5	A-Mem	GPT-4o-mini	495	0.692	✓
T5	A-Mem	Claude Haiku	495	0.738	✓
T5	A-Mem	Claude Sonnet	495	0.776	✓
T5	structured timeline	DeepSeek	495	0.748	✓
T5	structured timeline	GPT-4o-mini	495	0.776	✓
T5	structured timeline	Claude Haiku	495	0.869	✓
T5	structured timeline	Claude Sonnet	495	0.860	✓
T5	dense retrieval	DeepSeek	495	0.869	✓
T5	dense retrieval	GPT-4o-mini	495	0.888	✓
T5	dense retrieval	Claude Haiku	495	0.860	✓
T5	dense retrieval	Claude Sonnet	495	0.841	✓
T5	full context	DeepSeek	495	0.860	✓
T5	full context	GPT-4o-mini	495	0.897	✓
T5	full context	Claude Haiku	495	0.850	✓
T5	full context	Claude Sonnet	495	0.841	✓
T6	no-context (blind)	DeepSeek	700	0.358	✓
T6	no-context (blind)	GPT-4o-mini	700	0.311	✓
T6	no-context (blind)	Claude Haiku	700	0.000	✓
T6	no-context (blind)	Claude Sonnet	700	0.099	✓
T6	last-visit only	DeepSeek	700	0.424	✓
T6	last-visit only	GPT-4o-mini	700	0.470	✓
T6	last-visit only	Claude Haiku	700	0.470	✓
T6	last-visit only	Claude Sonnet	700	0.424	✓
T6	LLM summary	DeepSeek	700	0.444	✓
T6	LLM summary	GPT-4o-mini	700	0.391	✓
T6	LLM summary	Claude Haiku	700	0.417	✓
T6	LLM summary	Claude Sonnet	700	0.424	✓
T6	Mem0	DeepSeek	700	0.457	✓
T6	Mem0	GPT-4o-mini	700	0.424	✓
T6	Mem0	Claude Haiku	700	0.450	✓
T6	Mem0	Claude Sonnet	700	0.430	✓
T6	A-Mem	DeepSeek	700	0.503	✓
T6	A-Mem	GPT-4o-mini	700	0.464	✓
T6	A-Mem	Claude Haiku	700	0.444	✓
T6	A-Mem	Claude Sonnet	700	0.477	✓
T6	structured timeline	DeepSeek	700	0.609	✓
T6	structured timeline	GPT-4o-mini	700	0.530	✓
T6	structured timeline	Claude Haiku	700	0.523	✓
T6	structured timeline	Claude Sonnet	700	0.563	✓
T6	dense retrieval	DeepSeek	700	0.470	✓

(continued on next page)

(continued from previous page)

Task	Strategy	Backbone	n	Accuracy	Subset
T6	dense retrieval	GPT-4o-mini	700	0.391	✓
T6	dense retrieval	Claude Haiku	700	0.437	✓
T6	dense retrieval	Claude Sonnet	700	0.457	✓
T6	full context	DeepSeek	700	0.543	✓
T6	full context	GPT-4o-mini	700	0.497	✓
T6	full context	Claude Haiku	700	0.510	✓
T6	full context	Claude Sonnet	700	0.523	✓
T7	no-context (blind)	DeepSeek	400	0.256	✓
T7	no-context (blind)	GPT-4o-mini	400	0.244	✓
T7	no-context (blind)	Claude Haiku	400	0.012	✓
T7	no-context (blind)	Claude Sonnet	400	0.198	✓
T7	last-visit only	DeepSeek	400	0.360	✓
T7	last-visit only	GPT-4o-mini	400	0.302	✓
T7	last-visit only	Claude Haiku	400	0.140	✓
T7	last-visit only	Claude Sonnet	400	0.372	✓
T7	LLM summary	DeepSeek	400	0.512	✓
T7	LLM summary	GPT-4o-mini	400	0.477	✓
T7	LLM summary	Claude Haiku	400	0.349	✓
T7	LLM summary	Claude Sonnet	400	0.453	✓
T7	Mem0	DeepSeek	400	0.733	✓
T7	Mem0	GPT-4o-mini	400	0.640	✓
T7	Mem0	Claude Haiku	400	0.535	✓
T7	Mem0	Claude Sonnet	400	0.663	✓
T7	A-Mem	DeepSeek	400	0.651	✓
T7	A-Mem	GPT-4o-mini	400	0.616	✓
T7	A-Mem	Claude Haiku	400	0.442	✓
T7	A-Mem	Claude Sonnet	400	0.616	✓
T7	structured timeline	DeepSeek	400	0.640	✓
T7	structured timeline	GPT-4o-mini	400	0.616	✓
T7	structured timeline	Claude Haiku	400	0.512	✓
T7	structured timeline	Claude Sonnet	400	0.686	✓
T7	dense retrieval	DeepSeek	400	0.895	✓
T7	dense retrieval	GPT-4o-mini	400	0.860	✓
T7	dense retrieval	Claude Haiku	400	0.942	✓
T7	dense retrieval	Claude Sonnet	400	0.930	✓
T7	full context	DeepSeek	400	0.942	✓
T7	full context	GPT-4o-mini	400	0.872	✓
T7	full context	Claude Haiku	400	0.965	✓
T7	full context	Claude Sonnet	400	0.953	✓
T8	no-context (blind)	DeepSeek	134	0.231	✓
T8	no-context (blind)	GPT-4o-mini	134	0.254	✓
T8	no-context (blind)	Claude Haiku	134	0.172	✓
T8	no-context (blind)	Claude Sonnet	134	0.261	✓
T8	last-visit only	DeepSeek	134	0.313	✓
T8	last-visit only	GPT-4o-mini	134	0.351	✓
T8	last-visit only	Claude Haiku	134	0.179	✓
T8	last-visit only	Claude Sonnet	134	0.231	✓

(continued on next page)

(continued from previous page)

Task	Strategy	Backbone	n	Accuracy	Subset
T8	LLM summary	DeepSeek	134	0.231	✓
T8	LLM summary	GPT-4o-mini	134	0.254	✓
T8	LLM summary	Claude Haiku	134	0.194	✓
T8	LLM summary	Claude Sonnet	134	0.216	✓
T8	Mem0	DeepSeek	134	0.373	✓
T8	Mem0	GPT-4o-mini	134	0.358	✓
T8	Mem0	Claude Haiku	134	0.306	✓
T8	Mem0	Claude Sonnet	134	0.299	✓
T8	A-Mem	DeepSeek	134	0.373	✓
T8	A-Mem	GPT-4o-mini	134	0.336	✓
T8	A-Mem	Claude Haiku	134	0.291	✓
T8	A-Mem	Claude Sonnet	134	0.373	✓
T8	structured timeline	DeepSeek	134	0.448	✓
T8	structured timeline	GPT-4o-mini	134	0.396	✓
T8	structured timeline	Claude Haiku	134	0.537	✓
T8	structured timeline	Claude Sonnet	134	0.493	✓
T8	dense retrieval	DeepSeek	134	0.388	✓
T8	dense retrieval	GPT-4o-mini	134	0.276	✓
T8	dense retrieval	Claude Haiku	134	0.403	✓
T8	dense retrieval	Claude Sonnet	134	0.396	✓
T8	full context	DeepSeek	134	0.440	✓
T8	full context	GPT-4o-mini	134	0.269	✓
T8	full context	Claude Haiku	134	0.463	✓
T8	full context	Claude Sonnet	134	0.470	✓
T9	no-context (blind)	DeepSeek	1500	0.355	✓
T9	no-context (blind)	GPT-4o-mini	1500	0.685	✓
T9	no-context (blind)	Claude Haiku	1500	0.000	✓
T9	no-context (blind)	Claude Sonnet	1500	0.204	✓
T9	last-visit only	DeepSeek	1500	0.883	✓
T9	last-visit only	GPT-4o-mini	1500	0.895	✓
T9	last-visit only	Claude Haiku	1500	0.994	✓
T9	last-visit only	Claude Sonnet	1500	0.926	✓
T9	LLM summary	DeepSeek	1500	0.796	✓
T9	LLM summary	GPT-4o-mini	1500	0.775	✓
T9	LLM summary	Claude Haiku	1500	0.941	✓
T9	LLM summary	Claude Sonnet	1500	0.744	✓
T9	Mem0	DeepSeek	1500	0.806	✓
T9	Mem0	GPT-4o-mini	1500	0.818	✓
T9	Mem0	Claude Haiku	1500	0.941	✓
T9	Mem0	Claude Sonnet	1500	0.725	✓
T9	A-Mem	DeepSeek	1500	0.818	✓
T9	A-Mem	GPT-4o-mini	1500	0.784	✓
T9	A-Mem	Claude Haiku	1500	0.966	✓
T9	A-Mem	Claude Sonnet	1500	0.731	✓
T9	structured timeline	DeepSeek	1500	0.556	✓
T9	structured timeline	GPT-4o-mini	1500	0.664	✓
T9	structured timeline	Claude Haiku	1500	0.481	✓

(continued on next page)

(continued from previous page)

Task	Strategy	Backbone	n	Accuracy	Subset
T9	structured timeline	Claude Sonnet	1500	0.438	✓
T9	dense retrieval	DeepSeek	1500	0.772	✓
T9	dense retrieval	GPT-4o-mini	1500	0.710	✓
T9	dense retrieval	Claude Haiku	1500	0.920	✓
T9	dense retrieval	Claude Sonnet	1500	0.790	✓
T9	full context	DeepSeek	1500	0.701	✓
T9	full context	GPT-4o-mini	1500	0.698	✓
T9	full context	Claude Haiku	1500	0.907	✓
T9	full context	Claude Sonnet	1500	0.747	✓