

Who Is Asking? Personalized and Tiered Memory for Long-Lived AI Assistants

Yueqian Lin^{1,2} Sathiya Kumaran Mani² Yiran Chen¹
Sujata Banerjee² Vaishnavi Ranganathan²

¹Duke University ²Microsoft Research
y.lin@duke.edu vnattar@microsoft.com

Abstract

Personalized AI assistants in collaborative settings must balance private user context with shared team knowledge; i.e. the same question can have different answers, depending on who is asking. However, retrieving and reasoning over a growing shared archive, as agents do today, increases latency and cost. We present **GOLDILOCKS**, a multimodal assistive-agent demo that explores a tiered personal memory architecture which combines a lightweight private working *resident memory* alongside a shared *team memory*, both grounded in a common *eventstore*. A sufficiency gate first tries to answer from the low-cost resident tier and only escalates other tiers when needed. To evaluate long-lived personalized collaborative assistants, we also introduce **WHOQA**, a 500-question evaluation set built from public GitHub history, where work-related questions and answers trace to real pull requests and commits. On **WHOQA**, **GOLDILOCKS**'s deployed small-model configuration beats a strong hybrid RAG baseline by 7.0 points (59.0 vs. 52.0) at 13% lower estimated energy (a tier-weighted token cost), and reaches 62.4 (+10.4) with the answer model held fixed. The system serves 42.6% of queries from the resident tier. A compact social-memory scoping further lifts accuracy on teammate-specific personal details from 53 to 80. While we demonstrate with a set tier and scoping configuration, the system is generalizable to multiple tiers depending on the use case. The live demo lets users explore memory tiers, gating decisions, and interactively compare personalized responses across team personas.¹

1 Introduction

Today's meeting assistants answer a team's questions by searching the raw logs of every past session, with at most a user-specific memory on top.

¹Live demo: <https://goldilocks.yueqianlin.com/>.
Demo video: <https://youtu.be/Xs1f079wkGE>.

As the history accumulates, each query must retrieve its answer from an increasingly large and noisy archive. As latency, cost, and energy grow with the tokens read, delivering accuracy translates to paying more for every answer. Yet much of this knowledge is shared across teammates and repeatedly re-derived for each user while the rest is asker-specific. Per-user memory alone does not resolve this tension: storing everything per user duplicates shared content, while filtering shared memory by user can remove the cross-person evidence needed to answer some questions.

We present **GOLDILOCKS**, a personalized memory system with a live browser application, that separates memory by *scope* and reduces duplicates. **GOLDILOCKS** stores team knowledge into one shared tier (stored once), and keeps each persona's small, stable personal slice resident. A sufficiency gate then answers from the cheapest tier, using the smallest needed model, optimizing for *good enough*, and escalates to other tiers as needed. This helps accuracy while balancing cost and energy, where energy is estimated as a tier-weighted token cost. Scoping preserves the answers that are asker-specific, and a condensed index of personal details surfaces the sparse social facts that raw retrieval misses.

To evaluate personalized tiered memory, we need a *longitudinal*, multi-party corpus with reliable ground truth, and this is hard to obtain. Public meeting corpora lack longitudinal per-decision labels, and long-horizon dialogue benchmarks follow a single user. We therefore leverage existing public GitHub code records with rich history. The companion generator turns a set of related GitHub repositories into realistic, synthetic, weekly team meetings grounded in their pull-request and commit history. **GOLDILOCKS** ingests the meetings through an automatic speech recognition (ASR) and speaker identification block, emulating a deployed voice-based assistive agent (Lin et al.,

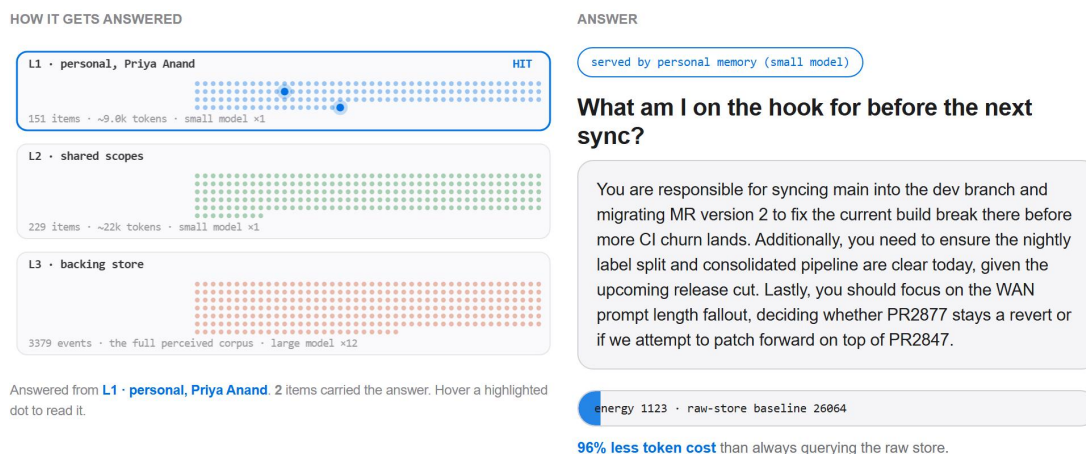


Figure 1: The live demo’s Q&A view. The gate probes the tech lead’s resident tier (L1), scores a HIT, and the small model answers from it. A miss falls through to the shared scopes and the backing store.

2025b, 2026a). Gold answers are derived directly from real PRs, commits, and decisions, rather than model-generated dialogue, yielding ground truth *by construction*. WHOQA is built on this corpus, and the generator is a reusable pipeline for scaling longitudinal data.

Among prior work in the memory domain, MemGPT and MemoryOS (Packer et al., 2023; Kang et al., 2025) page a *single* user’s history across stores, Mem0 and A-Mem (Chhikara et al., 2025; Xu et al., 2025) distill it into facts, and personalized retrieval conditions on a private profile (Zerhoubi and Granitzer, 2024). GOLDBLOCKS instead stores team-common content once in a *shared* tier, gives each persona a small resident tier lens over the same grounded corpus, and when this suffices it skips the backing store, jointly optimizing for accuracy and energy. Similarly, in the benchmark domain, long-horizon memory benchmarks LoCoMo (Maharana et al., 2024), Long-MemEval (Wu et al., 2025), and AlpsBench (Xiao et al., 2026) probe recall over a single user’s dialogues, whereas WHOQA makes the correct answer depend on *who is asking* over the same shared evidence, precisely the setting a deployed team assistant faces every day.

- **A live personalized and tiered-memory system:** GOLDBLOCKS stores team-common content once, keeps a resident memory per person, and a *sufficiency* gate attempts answers from the resident tier before escalating, accepting an answer only when its context is grounded, cites verbatim support, and covers the whole question.

- **A scalable corpus generator and WHOQA:** A data generation tool turns public repositories into a longitudinal meeting corpus grounded in real

history by construction, and WHOQA is the evaluation set built on it.

- **Accuracy and estimated-energy improvement:** On WHOQA, GOLDBLOCKS beats hybrid-RAG by +7.0 accuracy at 13% lower weighted token cost, our energy proxy (+10.4 with the answerer fixed), and the scoping gain transfers to other benchmarks.

2 Demonstration

GOLDBLOCKS is an application with a corpus generator: pointed at any set of related public GitHub repositories, the generator crawls their pull requests, issues, and commits, derives a weekly timeline from their history, and builds the corpus automatically. The hosted demo preprocesses two repositories as its running example. GOLDBLOCKS displays the three personas, each cast from a top repository contributor and given a carefully authored personality and personal-life backstory, the weekly meeting timeline, and the perceived conversation with transcripts behind every answer. The user then poses questions and sees the answer, the tier and model that served it, and the retrieved turns it was built from, annotated with context-token count and energy against an always-query-the-store baseline (Figure 1). A memory explorer tab replays the tiers week by week, visualizing how items enter, promote, and fall back over time.

What a demo visitor does. A visitor selects a persona and a cutoff week and asks a question in the first person, for example “what am I on the hook for before the next sync?”. The gate probes the persona’s personal memory, its cheap *resident tier*, and either answers from it with the small model, shown as a *hit* with its cost, or escalates to the shared team memory and the backing store on a

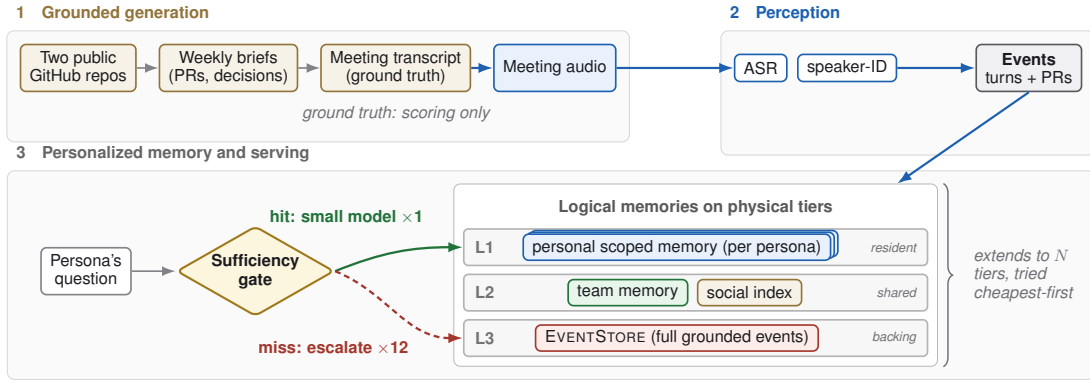


Figure 2: GOLDILOCKS overview. Two public repositories become audio-grounded meetings, and ASR with speaker-ID recovers the meeting turns from the audio alone (the ground-truth transcript is used only for scoring). These turns, plus code events from the repository record, feed the logical memories, each placed on the physical tier (L1–L3) that matches its access pattern and served cheapest-first by the sufficiency gate.

miss. The sharpest interaction is contrastive: re-asking the *same* question as a different persona returns a different correct answer from the *same* shared history, which asker-agnostic retrieval cannot produce. Each answer is annotated with its serving tier, context-token count, and energy, so the difference between a resident hit and a full-store escalation is visible at a glance: a hit answers from a few hundred retrieved tokens of the tech lead’s resident tier at a small fraction of the cost of a full-store escalation with the large model (Figure 1).

3 Logical and Physical Memory Tiers

GOLDILOCKS explores organizing memory along both logical and physical tiers (Figure 2). The logical tiers group memory by context, while the physical tiers organize memory by cost and frequency of access. Each logical tier is purpose-defined and consists of *scoped memories*, compact representations that distinguish between personal and team contexts; a *social-memory index*, which keeps personal facts retrievable; and a backing EVENTSTORE, which holds the full grounded-event corpus. Content is classified automatically when memories are condensed (Eq. 1 extracts personal-life facts). The physical tiers in this demo range from a per-persona *resident* tier (L1) to a shared *team* tier (L2) and a backing-store tier (L3), with access cost increasing from L1 to L3. The two hierarchies work in tandem to minimize duplication while keeping the most-consulted memories in the lowest-cost tier. Each logical tier can be further partitioned across physical tiers according to observed access frequency. A *sufficiency gate* serves each query across these tiers cheapest-first, so the better a persona’s memory predicts their relevant slice, the more questions the small model answers and the less energy the system spends on serving

the team. The architecture is not restricted to these tiers: additional project- or topic-specific scopes can be introduced as needed. The rest of the section describes the logical tiers in detail.

The EVENTSTORE: The hierarchy rests on a grounded-event corpus shared by all personas. Each item, a speaker-attributed meeting turn or associated code change, is tagged with its repository, meeting group, speaker, and week, and indexed for hybrid sparse and dense retrieval (Robertson and Zaragoza, 2009; Karpukhin et al., 2020). The store is the source of detailed truth, and the scoped memories never fork it. It occupies the physical backing-store tier (L3): large, complete, and consulted only when the cheaper tiers do not suffice.

SCOPED MEMORIES: Above the store sit the scoped memories, each a compact lens over it. A persona’s own meeting turns and authored pull requests up to the cutoff week, the “as of” point a demo user selects (Section 2), are condensed into a standing digest of the workstreams, files, and collaborators that persona touches, and updated incrementally as new meetings arrive. This personal digest is the resident tier (L1). Applying the same construction over all personas’ turns yields the shared team memory (L2), so what the team has in common is condensed once rather than copied into every personal digest. Two properties keep a lens correct when the answer depends on who is asking: First, each persona’s attributes and decisions stay distinct rather than merging into a team consensus (a merged summary gives everyone the same hedged reply). Second, personal-life facts stay separately addressable, extracted into a dedicated social index (below).

SOCIAL-MEMORY INDEX: The digests above are work-focused, and personal-life details (a team-mate’s trip, a pet, a home project) get buried in them: the detail is semantically far from the question, so raw-turn retrieval misses the gold turn most of the time. This is where a dedicated logical tier earns its place, extracting those facts into their own index. For each meeting m up to the cutoff, an extractor ϕ maps its turns to a set of (*subject*, *fact*) pairs, and the social-memory index is their union,

$$\mathcal{I} = \bigcup_m \phi(m), \quad \phi(m) = \{(p, a)\}, \quad (1)$$

where each pair names a persona p and records a fact a about them, for example (PRIYA, “adopted a rescue dog”). A social question about persona p is then answered from that persona’s slice of the index, written $\mathcal{I}(p) = \{a : (p, a) \in \mathcal{I}\}$, rather than from raw-turn retrieval, and the index lives beside the team memory in the shared tier (L2). Because it is shared, asker-dependent needles still rely on each persona’s personal tier, and Section 6 quantifies both effects.

SERVING AND ADMISSION: The sufficiency gate ties these tiers together. At query time it first tries the persona’s resident tier with the small model, and takes that answer only when it is grounded and sufficient: the model declares the context enough, cites a verbatim support span, and covers every part of the question (a *hit*, served from a few hundred tokens). Otherwise the query *escallates* to the team memory, still with the small model, then to the backing store with a larger model. Every tier is probed the same way, by sparse BM25 and dense rankings fused with reciprocal rank, so tiers differ only in what they hold. The priority in this design is to enable low-cost personalization while reducing information redundancy. Hence personalized abstraction is present in the resident tier. Over time, what stays in this tier is set by a utility-aware admission policy that heats items on useful retrieval, evicts cold item references to the store, and promotes items useful to several personas to the shared tier (constants as shown in Appendix B). This process can be visualized as a replay in the demo through the memory explorer feature. While the demo showcases a single EVENTSTORE for simplicity, other use cases can warrant separate backing stores for information in each persona’s stashed logic.

System	single-fact	persona-contr.	personal-recall	self-prep	pers.-work-bridge	Avg.
Hybrid-RAG	85.0	22.0	54.0	59.0	40.0	52.0
Per-user RAG	87.0	29.0	25.0	60.0	31.0	46.4
Mem0 (Chhikara et al., 2025)	82.0	26.0	67.0	67.0	29.0	54.2
GOLDILOCKS (Gemma-4-31B)	91.0	37.0	77.0	58.0	32.0	59.0
GOLDILOCKS (gpt-5.4)	90.0	39.0	80.0	61.0	42.0	62.4
Oracle	87.0	33.0	80.0	67.0	48.0	63.0

Table 1: Accuracy (%) on WHOQA.

4 Longitudinal Corpus Generator

A corpus for personalized memory must be longitudinal, and such data is scarce (Section 1). Development history fills the gap: related repositories record who did what, when, and why across months of pull requests, issues, and commits, so a corpus grounded in them evolves with time the way a real team’s history does. For each repository the generator retrieves the pull requests and issues (or commits) created in a time window, together with bodies, diffs, and review comments for concrete technical substance. Because the meetings are grounded in this deterministic record, the corpus carries verifiable ground truth.

From this record the generator casts its speakers. It intersects the repositories’ contributors, ranks them by combined activity, and casts the top three as the *lead*, *senior*, and *engineer* personas. Shared contributors are chosen deliberately: each persona’s work spans the repositories, so their memories are cross-repository and per-persona personalization has real content. Real accounts are mapped to fictional personas, and we never attribute synthesized dialogue to a named individual.

Meetings are generated out of the week-by-week record. Each week’s activity is folded into a *brief* (attendees, per-speaker updates, decisions, open items) that serves as the ground-truth scaffold, and gpt-5.4 expands the brief into a multi-party realistic transcript which includes personal anecdotes for the personas. A further pass renders it in spoken style, and a text-to-speech stage synthesizes it into realistic meeting audio with Higgs Audio v3 (Appendices B and C). From the briefs and transcripts the generator deterministically derives QA over four categories, from single-persona recall to cross-session synthesis (definitions in Appendix B). Because answers come from the scaffold rather than the model-written dialogue, each item carries a gold answer and its supporting turns. These scaffold QA are built into the generator, and WHOQA is a separate 500-question personalization evaluation set built on this corpus.

Everything above runs from one command. It takes the repository names, builds the corpus end

to end, and writes a run manifest recording the exact inputs and settings, so any corpus is auditable and re-derivable. The evaluated meeting turns are recovered from this synthesized audio rather than the clean transcript, and code events come from the repository record, so the corpus carries real perception noise. Audio synthesis runs on a local GPU or a hosted API, and a text-only mode can skip it at some cost to realism. The fetch-to-QA core needs only the GitHub CLI, so anyone can build the corpus from public activity and scale it to more repositories (Appendix B).

5 Evaluation Design

We evaluate GOLDILOCKS on WHOQA along the two axes a deployed assistant is judged on: accuracy and an energy proxy (the cost of answering a query, estimated from tokens rather than measured). GOLDILOCKS runs against a *long-context* baseline that places a persona’s whole history in the prompt and a *strong hybrid-RAG* baseline over the store, which instantiates the deployed practice of Section 1: retrieval over the raw record, with *per-user RAG* as its personalized variant (Table 1). Every tier and baseline uses the same hybrid retrieval, sparse BM25 (Robertson and Zaragoza, 2009) and dense bge-small-en-v1.5 (Xiao et al., 2024b) rankings fused by reciprocal rank, so configurations differ only in the memory they answer from. Each query yields the answer, the tier that served it, whether the gold evidence was retrieved, and the cost. A large language model (LLM) judge scores answers against the gold (Section 6). The energy proxy is a tier-weighted token cost: for a query q that reads $n_s(q)$ resident-tier and $n_\ell(q)$ escalation tokens,

$$E(q) = w_s n_s(q) + w_\ell n_\ell(q), \quad w_s=1, w_\ell=12, \quad (2)$$

reported as the mean over the run so that building and updating memory is separated from query-time cost, following a recent agent-memory characterization technique (Omri et al., 2026). The $w_\ell=12$ default is a conservative large-to-small per-token cost ratio (Poddar et al., 2025) (weight-sensitivity analysis in Appendix B). As a running example, the generator builds a corpus from vllm and vllm-omni, two related open-source LLM-serving repositories (24 meetings over twelve weeks, 2,995 turns), probed with WHOQA.

6 Results

Benchmark construction. WHOQA’s five categories span the two axes of Section 3: *single-fact* and *self-prep* (self-scoped work recall), *persona-contrastive* needles (one question, a different correct answer per persona, e.g., “*which revert am I on the hook for?*” points to a different PR for each asker), *personal recall* of a teammate’s social detail, and *personal-work bridge* across the content axis. The benchmark is balanced at 500 questions, 100 per category, each grounded in the persona’s pre-cutoff memory, with a leakage gate that drops any question whose supporting evidence post-dates its cutoff week (Appendix B). The deployed default answers gate hits with a small model (Gemma-4-31B). To isolate *which* memory each persona is served, the main comparison instead holds the answer model fixed at gpt-5.4 across configurations, sharing one prompt and judge, and scores are *judged accuracy* against the gold answer.

Table 1 compares configurations over the shared retriever of Section 5. The oracle routes each question to its gold tier with the same answerer, a reference rather than a strict ceiling. GOLDILOCKS’s deployed default, the small Gemma-4-31B answerer with large-model escalation on misses, beats hybrid-RAG by +7.0 accuracy at lower weighted cost, and the frontier answerer forms a +10.4 upper bound that leads every category against both RAG baselines and comes within 0.6 of the oracle. This suggests the residual gap lies mainly in retrieval rather than routing. The baselines fail differently: *Mem0*’s one shared fact set cannot vary by asker and stays at 26 on the contrastive needles, while *per-user RAG*, the per-user filtering alternative of Section 1, does personalize by filtering yet collapses on a teammate’s personal details because the filter discards the very evidence those questions need.

Analysis. The gain comes from two mechanisms on two kinds of question, not from routing, which already tracks the gold-evidence oracle within 0.6 points (above). *Scoping* wins the asker-dependent questions: on the persona-contrastive needles, where one question has a different correct answer per asker, GOLDILOCKS scores 39 against hybrid-RAG’s 22 and Mem0’s 26, because a single shared store answers everyone alike. *Abstraction* recovers what raw retrieval buries: a teammate’s personal detail sits far from the query in work-heavy memory, so its gold turn surfaces only 25% of the time at

System	Context tokens/q	Weighted cost/q	Small-tier %
Long-context	155,238	353×	0
Hybrid-RAG	440	1.00×	0
GOLDILOCKS	839	0.87×	42.6

Table 2: Query-time cost over 500 queries: context tokens and tier-weighted cost (small $\times 1$, large $\times 12$) relative to hybrid-RAG.

$k=12$, and extracting the shared personal facts into the social index lifts that to 76% and personal-recall accuracy from 53 to 80. That abstraction is worth +7.2 of the +10.4 gap over hybrid-RAG (scoping supplies the other +3.2), almost all of it in personal recall (5.4) and the contrastive needles (1.8). What remains is cross-axis retrieval: the two categories GOLDILOCKS scores lowest on, contrastive needles (39) and personal-work bridge (42), both join evidence across the scope and content axes, the clearest headroom.

Energy proxy. Serving 42.6% of queries from the small resident tier, GOLDILOCKS cuts weighted cost 13% below hybrid-RAG and 99.8% below long-context (Eq. 2, Table 2). To be explicit about what this proxy measures: GOLDILOCKS reads *more* raw context tokens per query than hybrid-RAG (839 vs. 440, Table 2); the saving arises because most of those tokens are charged at the resident-tier weight $w_s=1$ rather than the $w_\ell=12$ escalation weight, and we report token accounting rather than measured latency or joules. The exact weight is not load-bearing: the saving holds for any large-to-small ratio above about $5\times$, and the $w_\ell=12$ default is conservative (Appendix B).

How many tiers are needed? The Goldilocks N . The number of probe stops N the gate tries cheapest-first is a knob, where each stop is one memory the gate can probe, whether a scoped memory, the social index, or the store, and this ablation shows how to set it for a corpus. We vary N from 1 (the flat store) to 5, adding the *personal* memory, *team* memory, *social* index, and a routed *project/topic* scope as stops, on a balanced 250-question subset of WHOQA with the same gate, answerer, and judge. Accuracy plateaus after the first resident tier while weighted cost keeps falling to a minimum at $N^*=4$, the deployed configuration (Figure 3). A fifth tier adds less than one point while raising cost, so the value lies in scopes that reflect user differences.

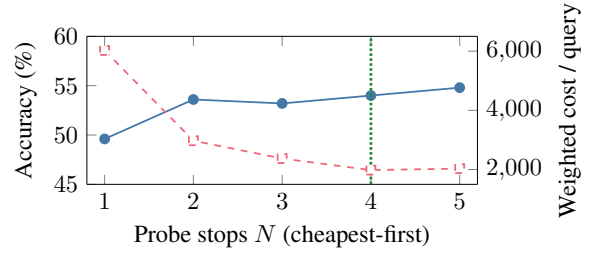


Figure 3: Accuracy (blue, left) and weighted cost (red dashed, right) versus probe stops N , with the cost minimum at $N^*=4$ (dotted).

System	n	Acc.	Reader tok/q
<i>LoCoMo</i> (judged accuracy, %)			
Flat (hybrid-RAG)	195	29.7	—
GOLDILOCKS (scoped)	195	44.1	—
<i>AlpsBench</i> (Task-2 top-1 accuracy)			
Flat RAG	361	0.25	0
Flat + LLM reads all [§]	60	0.80	~12k
GOLDILOCKS (scoped)	361	0.73	0
GOLDILOCKS scoped + reader [§]	60	0.97	~30

Table 3: External validation on **LoCoMo** (Maharana et al., 2024) and **AlpsBench** (Xiao et al., 2026): scoping to the asker beats flat retrieval on both, and a small reader over the asker’s tier matches full-pool reading at far fewer tokens (setup in Appendix B). [§]Gemma-4-31B reader, 60-query sample.

6.1 External validation

To test whether the gain generalizes beyond our generated corpus, we validate GOLDILOCKS on two independent, published benchmarks it was not designed for. On **LoCoMo** (Maharana et al., 2024), a long-term dialogue memory benchmark, scoping the same pooled store to the asker lifts judged accuracy by +14.4 points and holds in 9 of 10 dialogues (Table 3). An adapter maps sessions to weeks and turns to speaker-attributed events, each question asked by the speaker who related the fact (leakage-checked, adversarial category excluded). LoCoMo is not persona-contrastive, so this isolates scoped versus flat retrieval.

On **AlpsBench** (Xiao et al., 2026), a personal-memory benchmark that pools all users’ memories from 2,500 WildChat sessions into one flat store, we test whether scoped retrieval survives store growth. Flat retrieval degrades as the pool grows while scoping to the asker holds: at 1,276 memories flat top-1 falls from 0.54 to 0.25 while the asker’s tier holds at 0.73 (361 Task-2 queries, Figure 4), and a small Gemma-4-31B reader over the asker’s ~ 3 -candidate tier reaches 0.97 from ~ 30 tokens, against 0.80 from $\sim 12k$ tokens reading the full pool (Table 3). The gap survives a tf-idf retriever (0.09 flat vs 0.51 scoped).

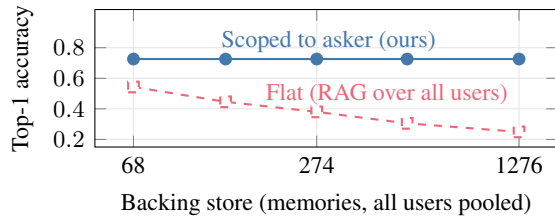


Figure 4: AlpsBench top-1 accuracy as the pooled store grows, for flat versus asker-scoped retrieval.

7 Conclusion

GOLDILOCKS shows a team assistant’s memory can know who is asking. Storing what a team shares once, keeping each member’s slice resident, and stopping at the smallest sufficient model improves accuracy and cuts cost at once. The demo makes these mechanics inspectable, and the assistant more personal and cheaper.

Limitations

GOLDILOCKS evaluates on synthetic meetings grounded in real pull requests, which yields verifiable gold answers that private recordings cannot offer, and the scoping mechanism transfers to Lo-CoMo and AlpsBench. The demo corpus covers two repositories from one organization, with model-written small talk supplying the personal-life material. The answerer and judge share one model family, and Appendix B confirms every ranking holds under a different-family judge, a judge-free metric, and a closed-book control. Because no baseline sees the same evidence as GOLDILOCKS, the reported gains combine asker-aware scoping with tiering and gating. Natural next steps include a matched flat baseline over the union of personal and shared evidence, and multi-user adaptations of the single-user systems in Tables 1 and 4. We treat WHOQA, built from the same weekly scaffolds, as an evaluation artifact accompanying the demo rather than a standalone benchmark. The prototype personalizes retrieval without yet enforcing access control over the shared social index, and the efficiency numbers report the token-based proxy of Eq. 2 rather than measured latency.

Ethics Statement

GOLDILOCKS uses only public repository data and synthesizes fictional meetings from it: contributors are mapped to fictional personas, and no synthesized dialogue is attributed to a real person. The speaker timbres are randomly sampled. No private communications or recordings are used at any step. The corpus design guarantees this by construction.

Towards any future open release, the corpora carry a provenance manifest and are re-derivable from it, so no third-party repository text is redistributed.

Acknowledgments

This work was supported in part by NSF 2112562 and ARO W911NF-23-2-0224.

References

- Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, and 9 others. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, PMLR 162.
- Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. [The AMI meeting corpus: A pre-announcement](#). In *Machine Learning for Multimodal Interaction (MLMI)*, pages 28–39.
- Tuochao Chen, Nicholas Batchelder, Alisa Liu, Noah A. Smith, and Shyamnath Gollakota. 2025. LlamaPIE: Proactive in-ear conversation assistants. In *Findings of the Association for Computational Linguistics: ACL 2025*.
- Yuhao Chen, Yi Xu, Xinyun Ding, Xiang Fang, Shuochen Liu, Luxi Lin, Qingyu Zhang, Ya Li, Quan Liu, and Tong Xu. 2026. VehicleMemBench: An executable benchmark for multi-user long-term memory in in-vehicle agents. *arXiv preprint arXiv:2603.23840*.
- Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. [Mem0: Building production-ready AI agents with scalable long-term memory](#). In *Proceedings of the 28th European Conference on Artificial Intelligence (ECAI)*, pages 2993–3000.
- Harshita Chopra, Krishna Kant Chintalapudi, Suman Nath, Ryen W. White, and Chirag Shah. 2026. Thinking ahead: Prospection-guided retrieval of memory with language models. *arXiv preprint arXiv:2605.14177*.
- Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. 2009. [Reciprocal rank fusion outperforms Condorcet and individual rank learning methods](#). In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 758–759.

- Yudai Honda, Yuki Fujita, Keiichi Zempo, and Shogo Fukushima. 2025. [Human-like remembering and forgetting in LLM agents: An ACT-R-inspired memory architecture](#). In *Proceedings of the 13th International Conference on Human-Agent Interaction (HAI)*, pages 229–237.
- Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesht, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. RULER: What’s the real context size of your long-context language models? In *First Conference on Language Modeling (COLM)*.
- Chuanrui Hu, Tong Li, Xingze Gao, Hongda Chen, Yi Bai, Dannong Xu, Tianwei Lin, Xiaohong Li, Yunyun Han, Jian Pei, and Yafeng Deng. 2026. Evaluating long-horizon memory for multi-party collaborative dialogues. *arXiv preprint arXiv:2602.01313*.
- Yebowen Hu, Tim Ganter, Hanieh Deilamsalehy, Franck Dernoncourt, Hassan Foroosh, and Fei Liu. 2023. [MeetingBank: A benchmark dataset for meeting summarization](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. 2024. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.
- Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. 2025. Memory OS of AI agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Andrej Karpathy. 2026. LLM wiki: A pattern for personal knowledge bases. GitHub gist, <https://gist.github.com/karpathy/442a6bf555914893e9891c11519de94f>.
- Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.
- Yueqian Lin, Yuzhe Fu, Jingyang Zhang, Yudong Liu, Jianyi Zhang, Jingwei Sun, Hai Li, and Yiran Chen. 2025a. SpeechPrune: Context-aware token pruning for speech information retrieval. In *IEEE International Conference on Multimedia and Expo (ICME)*.
- Yueqian Lin, Zhengmian Hu, Jayakumar Subramanian, Qinsi Wang, Nikos Vlassis, Hai Li, and Yiran Chen. 2025b. AsyncVoice agent: Real-time explanation for LLM planning and reasoning. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Demo Track*.
- Yueqian Lin, Zhengmian Hu, Qinsi Wang, Yudong Liu, Hengfan Zhang, Jayakumar Subramanian, Nikos Vlassis, Hai Li, and Yiran Chen. 2026a. Voice evaluation of reasoning ability: Diagnosing the modality-induced performance gap. In *I Can’t Believe It’s Not Better Workshop at the International Conference on Learning Representations (ICLR)*.
- Yueqian Lin, Jingyang Zhang, Qinsi Wang, Hancheng Ye, Yuzhe Fu, Yudong Liu, Hai Li, and Yiran Chen. 2026b. HippoMM: Hippocampal-inspired multi-modal memory for long audiovisual event understanding. In *Findings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Xunzhuo Liu, Bowei He, Xue Liu, Andy Luo, Haichen Zhang, and Huamin Chen. 2026. Knowledge access beats model size: Memory augmented routing for persistent AI agents. *arXiv preprint arXiv:2603.23013*.
- 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 (ACL)*, pages 13851–13870.
- Tony Mason. 2026. The missing memory hierarchy: Demand paging for LLM context windows. *arXiv preprint arXiv:2603.09023*.
- Yasmine Omri, Ziyu Gan, Zachary Broveak, Robin Geens, Zexue He, Alex Pentland, Marian Verhelst, Tsachy Weissman, and Thierry Tambe. 2026. Agent memory: Characterization and system implications of stateful long-horizon workloads. *arXiv preprint arXiv:2606.06448*.
- Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2023. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*.
- Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. [Generative agents: Interactive simulacra of human behavior](#). In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*.
- Soham Poddar, Paramita Koley, Janardan Misra, Niloy Ganguly, and Saptarshi Ghosh. 2025. Towards sustainable NLP: Insights from benchmarking inference energy in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*.
- Hongjin Qian, Zheng Liu, Peitian Zhang, Kelong Mao, Defu Lian, Zhicheng Dou, and Tiejun Huang. 2025. MemoRAG: Boosting long context processing with global memory-enhanced retrieval augmentation. In *Proceedings of the ACM Web Conference (WWW)*.

- Alireza Rezazadeh, Zichao Li, Ange Lou, Yuying Zhao, Wei Wei, and Yujia Bao. 2025. Collaborative memory: Multi-user memory sharing in LLM agents with dynamic access control. *arXiv preprint arXiv:2505.18279*.
- Stephen Robertson and Hugo Zaragoza. 2009. [The probabilistic relevance framework: BM25 and beyond](#). *Foundations and Trends in Information Retrieval*, 3(4):333–389.
- Alireza Salemi and Hamed Zamani. 2025. LaMP-QA: A benchmark for personalized long-form question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. [Neural theory-of-mind? on the limits of social intelligence in large LMs](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3762–3780.
- Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2025. Long-MemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations (ICLR)*.
- Menglin Xia, Xuchao Zhang, Shantanu Dixit, Paramaguru Harimurugan, Rujia Wang, Victor Ruhle, Robert Sim, Chetan Bansal, and Saravan Rajmohan. 2026. Memora: A harmonic memory representation balancing abstraction and specificity. *arXiv preprint arXiv:2602.03315*.
- Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024a. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*.
- Jianfei Xiao, Xiang Yu, Chengbing Wang, Wuqiang Zheng, Xinyu Lin, Kaining Liu, Hongxun Ding, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2026. AlpsBench: An LLM personalization benchmark for real-dialogue memorization and preference alignment. *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.
- Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muenighoff, Defu Lian, and Jian-Yun Nie. 2024b. [C-pack: Packed resources for general chinese embeddings](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.
- 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 38 (NeurIPS)*.
- Yifan Yu, Yu Gan, Nikhil Sarda, Lillian Tsai, Jiaming Shen, Yanqi Zhou, Arvind Krishnamurthy, Fan Lai, Henry M. Levy, and David Culler. 2025. [IC-Cache: Efficient large language model serving via in-context caching](#). In *Proceedings of the 31st ACM Symposium on Operating Systems Principles (SOSP)*, pages 375–398.
- Saber Zerhouni and Michael Granitzer. 2024. PersonaRAG: Enhancing retrieval-augmented generation systems with user-centric agents. In *Proceedings of the IR-RAG Workshop at SIGIR 2024 (CEUR Vol. 3784)*, pages 1–11.
- Haozhen Zhang, Haodong Yue, Tao Feng, Quanyu Long, Jianzhu Bao, Bowen Jin, Weizhi Zhang, Xiao Li, Jiaxuan You, Chengwei Qin, and Wenya Wang. 2026a. Learning query-aware budget-tier routing for runtime agent memory. In *Proceedings of the International Conference on Machine Learning (ICML)*.
- Hengfan Zhang, Yueqian Lin, Hai Li, and Yiran Chen. 2026b. Bridging the perception gap: A lightweight coarse-to-fine architecture for edge audio systems. In *Edge AI Research Symposium*.
- Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. 2023. H2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*.
- Wanjuan Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. MemoryBank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19724–19731.

System	Per-persona	Shared	Tiered	Gate sm→lg	Energy
Long-context	×	✓	×	×	×
Flat RAG (top- <i>k</i>)	×	✓	×	×	×
MemGPT / MemoryOS	×	×	✓	~	×
Collab. Memory	~	✓	×	×	×
GOLDILOCKS	✓	✓	✓	✓	✓

Table 4: GOLDILOCKS vs. representative memory/retrieval systems. *Per-persona*: answer may differ by asker. *Shared*: one multi-user corpus. *Tiered*: cheap resident memories under the store. *Gate*: small model answers and escalates on a miss. *Energy*: reports cost (~ partial).

A Related Work

Agent memory and hierarchies. LLM agents gain persistent memory through extraction, summarization, and retrieval (Mem0 (Chhikara et al., 2025), MemGPT/Letta (Packer et al., 2023), A-Mem (Xu et al., 2025), Generative Agents (Park et al., 2023)), and the view of context limits as a memory-hierarchy problem (Mason, 2026) motivates our tiering. The closest tiered system, MemoryOS (Kang et al., 2025), pages a *single* user’s memory across stores, and the closest retrieval mechanism, MemoRAG (Qian et al., 2025), drafts from a global memory to guide retrieval but still invokes the expensive reader on every query. Memora (Xia et al., 2026) balances abstraction against specificity within one memory, prospection-guided retrieval (Chopra et al., 2026) probes a user’s memory with imagined future queries, and HippoMM (Lin et al., 2026b) consolidates long audio-visual streams into episodic memory, all again for a single user. GOLDILOCKS differs on both: it scopes a *shared* corpus *per persona*, *skips* the store when a resident tier suffices, and reports energy. Like Karpathy’s *LLM Wiki* (Karpathy, 2026) it reuses a compiled memory rather than re-deriving by RAG.

Personalized and multi-user memory. Personalized retrieval augments queries with a user profile (Zerhoubi and Granitzer, 2024), and personalization matters measurably: a mismatched profile degrades answer quality by large margins (Salemi and Zamani, 2025). Most such work personalizes over each user’s *private* history. The multi-user setting over a *shared* corpus has been studied for in-vehicle assistants that identify who is speaking and apply that user’s stored preference (Chen et al., 2026) and for role-conditioned multi-party recall scored for accuracy alone (Hu et al., 2026). Closest is Collaborative Memory (Rezazadeh et al., 2025), which routes shared items by permission with provenance but does not adapt admission or re-

port efficiency. GOLDILOCKS adds a utility-aware admission policy (replayed in the demo) across persona and shared scopes, evaluates *contrastive* memory where one question over the same shared evidence has a different correct answer per persona, and scores accuracy alongside a model-free *energy* proxy. Serving a different answer per persona also requires a model of what each persona knows, a lightweight form of *theory-of-mind* evaluation (Sap et al., 2022).

Eviction and forgetting. We borrow KV-cache heavy-hitter and attention-sink intuitions (H2O (Zhang et al., 2023), StreamingLLM (Xiao et al., 2024a)) but apply them to *durable*, *cross-session* items, with a utility score that rises on use and decays over time in the spirit of MemoryBank’s forgetting curve (Zhong et al., 2024) and ACT-R-style decay (Honda et al., 2025). Unlike Generative Agents, which down-weight but never evict (Park et al., 2023), a forgotten item here falls to a cheaper tier rather than vanishing.

Cost-aware routing and retrieval. Adaptive retrieval decides *whether* to retrieve from query complexity (Jeong et al., 2024), and memory systems route across budget or capacity tiers (Zhang et al., 2026a; Liu et al., 2026). These route on *query* difficulty, whereas GOLDILOCKS routes on whether a *persona*’s scoped memory suffices and reports a query-time energy proxy (Omri et al., 2026). Unlike semantic or in-context caching that reuses past answers (Yu et al., 2025), GOLDILOCKS caches nothing: its saving comes from a small, well-scoped working set, echoing retrieval-substitutes-for-parameters (Borgeaud et al., 2022), token pruning over long speech contexts (Lin et al., 2025a), and small-then-large assistants that escalate on uncertainty (Chen et al., 2025; Zhang et al., 2026b).

Longitudinal and meeting evaluation. Long-MemEval (Wu et al., 2025) and LoCoMo (Maharana et al., 2024) probe long-horizon recall over constructed conversations, and AlpsBench (Xiao et al., 2026) filters real human-LLM dialogues into per-user memory tasks. WHOQA instead grounds multi-party *team* meetings in public development activity for per-decision, multi-persona ground truth. Real meeting corpora (AMI (Carletta et al., 2006), MeetingBank (Hu et al., 2023)) lack this longitudinal, per-decision labeling, and length-stress probes like RULER (Hsieh et al., 2024) hide a fact in filler, whereas our *persona-contrastive* setting hides the *personalized* answer across personas.

B Implementation details

Retrieval and indexing. The evaluated store indexes the ASR-perceived turns recovered from the synthesized audio, not the ground-truth transcript, so every reported number includes perception noise and the transcript serves only for scoring. Perception runs Whisper large-v3 for transcription and ECAPA voice embeddings for speaker attribution. Every store item is abstracted to a common text spine and indexed twice: sparse BM25 (Robertson and Zaragoza, 2009) and dense embeddings (Xiao et al., 2024b) from BAAI/bge-small-en-v1.5 (run locally via fastembed). The two rankings are fused with reciprocal rank fusion (Cormack et al., 2009) ($k=60$), and the gate’s tier probes read the top 10 fused turns per query, while store reads (hybrid-RAG, per-user RAG, and gate escalations) use the top 12.

Admission constants. The weekly heat update multiplies an item’s heat by a decay of $\lambda=0.5$ and adds a bump of +3 when the item is retrieved and used in a judge-supported answer that week, +1 when it is only retrieved, and 0 otherwise, following an Ebbinghaus-style forgetting curve (Zhong et al., 2024; Park et al., 2023). An unpinned item evicts to the store once its heat falls below 0.3 and promotes one tier once it helps at least two personas beyond its owner. Pinned items, a persona’s open obligations and personal dossier, are always retained.

Generator stages and caching. Repository activity is fetched with the GitHub CLI, recording each pull request or issue with its bodies, diffs, and review comments. Transcript expansion adds per-turn prosody, an oralization pass inserts natural disfluencies, and text-to-speech renders the result with Higgs Audio v3, one fixed voice per persona (Appendix C). Every stage is cached and idempotent, and the run manifest records the repositories, PR/commit ranges, login→persona map, model, seed, and output counts.

Scaffold QA categories. The generator’s built-in QA covers four categories: *per-persona retrieval* (what did person p work on in week w ?), *decision tracking*, *action-item follow-up* (cross-week consistency), and *cross-session synthesis*, all derived deterministically from the weekly briefs.

WHOQA construction. WHOQA holds 500 questions, 100 per category (definitions in Section 6). Every item is grounded in the persona’s

pre-cutoff memory, and a leakage gate drops any question whose evidence post-dates its cutoff.

Energy-weight sensitivity. The $\times 12$ default is a *conservative* large-to-small per-token cost ratio: measured LLM inference energy climbs steeply with model scale (Poddar et al., 2025), and current frontier-to-small serving prices differ by more than an order of magnitude, so 12 sits below typical ratios and if anything understates GOLDBLOCKS’s savings. Table 2 implies GOLDBLOCKS averages about 498 small-tier and 341 escalated context tokens per query, so its weighted cost relative to hybrid-RAG is $(498 + 341w)/(440w)$ at weight w . This crosses 1.0 near $w^*\approx 5$, equals the reported 0.87 at $w=12$, and falls toward a 0.78 floor as w grows.

Judge and contamination checks. A rerun of GOLDBLOCKS and hybrid-RAG scored by a different-family judge (Gemma-4-31B) preserves the ranking in every category (68.4 vs 56.4 average accuracy), and a deterministic key-term metric computed alongside the judge agrees (51.7 vs 45.0), so the comparison does not rest on the gpt-5.4 judge. A closed-book control bounds pretraining leakage: told the source repositories and free to use parametric knowledge, the same answer model reaches 8.8 average accuracy, far below every retrieval configuration.

Deployed small-model path. The default row of Table 1 answers the gate’s L1 and L2 hits (81% of queries; 42.6% from L1 alone) with Gemma-4-31B and escalates misses to gpt-5.4, under the same benchmark, prompts, and judge as every other row.

External-validation setup (Table 3). LoCoMo uses 195 questions with the gpt-5.4 judge; Alps-Bench uses the dev split, Task-2 ownership, the full 1,276-memory store, and the main runs’ embedder. Our retriever reproduces the released baseline (0.41 vs 0.40), the scoped-versus-flat gap is retriever-independent (tf-idf: 0.09 vs 0.51), and at the full store flat retrieval falls below a random pick among the asker’s own memories (0.36).

C Audio synthesis details

Complete audio by chunked re-synthesis. Higgs Audio v3 truncates long turns, voicing only the first roughly 80–120 words. Turns whose implied speaking rate exceeds 4.5 words per second are re-rendered as sentence-bounded chunks under seventy words and concatenated.