ADAR: Abductive Distance-Aware Retrieval for Implicit-Preference Dialogue Memory

Anonymous ACL submission

Abstract

Dialogue-memory benchmarks have shifted from explicit fact recall to implicit preference inference, where the relevant evidence is dispersed across many weakly suggestive turns and the next-step retrieval signal is structurally missing. Existing approaches falter on both ends of this trajectory: memory systems precompute query-blind indices, and search-augmented retrieval is designed for factual multi-hop QA over shared corpora. We introduce the **Abductive Distance Score** (ADS), a single bounded score that combines a lexical anchor-coverage distance with a contrastive semantic distance. ADS plays two complementary roles: as an offline diagnostic on gold evidence it provides a lens for inspecting how indirect each query in a dataset really is, and as an online statistic on the cumulative retrieval pool it gives an inference-time signal that an LLM can read while it searches. Building on ADS, we train **Abductive Distance-Aware Retrieval** (ADAR), a 4B dialogue search agent, through restricted-oracle SFT in which a 27B teacher with privileged gold access produces a candidate-trajectory pool, and filtering then keeps only the well-grounded trajectories that a 4B student is fine-tuned on. Across PersonaMem-v2, LoCoMo, and LongMemEval, which span the explicit-to-implicit spectrum, ADAR achieves a 7.32% average gain over SOTA baselines.[1]

1 Introduction

Personal assistants are increasingly expected to act on implicit user preferences rather than explicit statements, which users seldom declare in canonical form. Preferences leak out of routine task interactions, and the assistant must reuse scattered signals to give a personalised response. Recent dialogue-memory benchmarks track this trend by moving the ground-truth target from facts the user has stated (Maharana et al., 2024; Wu et al., 2025) to preferences the user has revealed (Zhao et al., 2025; Peng et al., 2025; Jiang et al., 2025a,b; Xiao et al., 2026; Uddin et al., 2026). Figure 1 makes this indirect style concrete: when a user asks "recommend an exercise for me" and the only relevant evidence in the history reads "doctor diagnosed user with heart disease", the two share no surface anchors, and the gold turn is not close to the query in the embedding space. There is no single explicit fact in the history that can be directly looked up, so both the relevance criterion and the retrieval target become non-trivial to specify in advance.

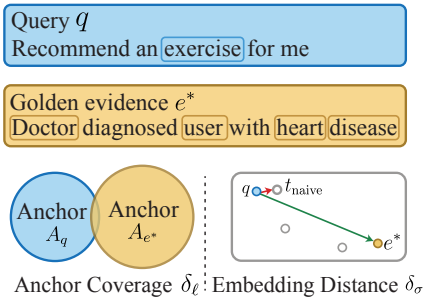

Figure 1: The two ADS axes on a PersonaMem-v2 example. Boxed tokens are anchors. **Left:** δ_ℓ is the IDF-weighted fraction of query anchors A_q not present in evidence anchors A_{e^*}. **Right:** the green arrow is $s_{\text{in}} = \cos(q, e^*)$ and the red is $s_{\text{out}} = \cos(q, t_{\text{naive}})$. Here $s_{\text{in}} \leq s_{\text{out}}$ so $\delta_\sigma \geq 1$ and the example is adversarial. The combined ADS is $\sqrt{\delta_\ell \cdot \delta_\sigma}$.

Two adjacent lines address pieces of this problem, but both inherit a "fact coverage" view of retrieval. **Memory systems** (Packer et al., 2024; Zhong et al., 2024; Xu et al., 2025; Chhikara et al., 2025; Fang et al., 2025; Wang et al., 2026) build structured user representations at ingestion time and rely on similarity search at query time, assuming the relevance criterion can be encoded into the index without seeing the query. **Search-augmented LLMs** (Gao et al., 2023; Trivedi et al., 2023; Asai et al., 2024; Jeong et al., 2024; Li et al.,

[1] Our code is publicly available at https://anonymous.4open.science/r/ADAR-844F

2025; Jin et al., 2025; Song et al., 2025; Zheng et al., 2025; Sun et al., 2025) decide retrieval actions at inference time, but are coupled to factual multi-hop QA over shared corpora and so chase explicit fact tokens rather than the dispersed cues a preference leaves behind. Neither line directly answers what makes an instance abductive, or how an agent should search when no such fact exists in the history in the first place.

We make this gap quantitative by introducing the **Abductive Distance Score** (ADS), a bounded statistic that measures how far a query is from its supporting evidence along two complementary axes: lexical anchor coverage and contrastive semantic separation. Applied to the gold-annotated evidence turn, ADS gives **offline ADS**, a difficulty diagnostic at both the benchmark and the instance level. Applied instead to the cumulative pool of turns retrieved so far, ADS gives **online ADS**, a label-free retrieval-saturation signal that decreases monotonically as the pool grows.

Building on this metric, we train **Abductive Distance-Aware Retrieval** (ADAR), a 4B-parameter dialogue search agent, through a non-RL pipeline. A frozen 27B teacher runs the agent loop with access to the gold evidence and answer but under explicit anti-leak constraints, and produces a large pool of candidate trajectories per query that demonstrate the per-round search query, the pool-saturation reading, and when to commit. Three filters for anti-leak, outcome, and evidence-presence keep only the trajectories that genuinely retrieve the gold turn before committing, and a Qwen3.5-4B student is fine-tuned on the survivors. The intuition behind this sample-and-filter pipeline is that, when the input does not carry an explicit next-step retrieval signal, the student needs a demonstration of how to look, not just a signal of whether it succeeded. A 1-bit outcome reward does not tell the agent what the target evidence looks like and therefore cannot guide the search direction at intermediate rounds, whereas a gold-aware teacher can demonstrate that direction step by step, and label-aware filtering removes candidates that happened to land on the right answer for the wrong reason.

We examine the method along three axes. On method effectiveness, across PersonaMem-v2, LoCoMo, and LongMemEval, three representative points of the explicit-to-implicit trajectory, ADAR outperforms every 4B-class training-free and RL-trained baseline. On the value of offline ADS, the seven evaluated benchmarks cleanly separate on a two-dimensional difficulty plane, providing a quantitative lens on query-evidence indirectness. On the value of online ADS, it is a signal an LLM can consult while it searches: the plateau height tracks a query's intrinsic indirectness and the per-round descent tracks retrieval progress.

Our main contributions are as follows:

- To quantify how far a query lies from its supporting evidence and diagnose difficulty, we propose the Abductive Distance Score (ADS).
- To teach an agent how to search under query-evidence gaps, we propose Abductive Distance-Aware Retrieval (ADAR), trained via a restricted-oracle SFT pipeline.
- Empirically, we demonstrate consistent improvements over SOTA baselines, spanning both training-free and RL-based search agents, with an average gain of 7.32%.

2 Related Work

2.1 Dialogue Memory Benchmarks

Dialogue memory benchmarks have evolved along an explicit-to-implicit axis. LoCoMo (Maharana et al., 2024) and LongMemEval (Wu et al., 2025) target long-horizon factual recall over multi-session conversations and define their ground truth from statements the user has explicitly made. Subsequent benchmarks shift the target to preferences the user reveals only behaviourally. PrefEval (Zhao et al., 2025) formalises preference following in 100k-token persona dialogues. IP-Dialog (Peng et al., 2025) generalises this to twelve categorical user attributes. PersonaMem-v1 (Jiang et al., 2025a) introduces long-horizon interactions with seven in-situ query types. PersonaMem-v2 (Jiang et al., 2025b) pushes the implicit setting further by embedding behavioural cues into routine task interactions, and the concurrent benchmark Alps-Bench (Xiao et al., 2026) extends the trajectory toward real conversation logs.

Across the trajectory, no benchmark provides a numerical measure of how indirect each instance is, so cross-benchmark comparison has lacked a difficulty-aware lens, resting on accuracy alone.

2.2 Dialogue Memory Methods

A dominant paradigm builds structured user representations during ingestion: MemGPT (Packer et al., 2024) swaps memory pages OS-style, MemoryBank (Zhong et al., 2024) models Ebbinghaus-style decay, A-MEM (Xu et al., 2025) links

Zettelkasten notes, Mem0 (Chhikara et al., 2025) consolidates facts into a graph index, LightMem (Fang et al., 2025) uses sleep-time consolidation, and MemMachine (Wang et al., 2026) preserves raw episodes for sentence-level retrieval. They differ in how memory is compressed but share a premise: the ingested representation should provide a complete set of handles such as facts, entities, or episodes that any future query can match by vector similarity, without knowing those queries in advance, an assumption abductive queries violate.

This works for explicit fact lookup but is structurally insufficient for implicit preference inference, whose relevance criterion is itself query-dependent. The user's preference dimensions are open-ended, and the truly relevant turn is often lexically and semantically far from the query (as in the exercise/heart-disease example of Figure 1), so ingestion-time consolidation cannot reliably mark which turns will later count as evidence.

2.3 Search-Augmented Large Language Models

A parallel line equips LLMs with active retrieval. Training-free methods include hypothetical-document rewriting (Gao et al., 2023), interleaved chain-of-thought and retrieval (Trivedi et al., 2023), reflection-token-controlled retrieval (Asai et al., 2024), complexity-routed retrieval (Jeong et al., 2024), and agentic search-and-reason loops with document grounding (Li et al., 2025). RL-trained search agents have rapidly followed: Search-R1 (Jin et al., 2025), R1-Searcher (Song et al., 2025), DeepResearcher (Zheng et al., 2025) and ZeroSearch (Sun et al., 2025) train an LLM to interleave search calls with reasoning under outcome-based rewards, all over shared factual corpora.

These methods all train and evaluate on factual QA over shared corpora (Wikipedia or the open web), spanning single-hop (Natural Questions (Kwiatkowski et al., 2019)) and multi-hop (HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (Ho et al., 2020), MuSiQue (Trivedi et al., 2022), Bamboogle (Press et al., 2023)) settings. There every answer is a fact, a string appearing verbatim in the corpus, and the retrieval policy is selected and rewarded by how well it covers such facts. This template suits retrieving and extracting a known fact, or a chain of explicitly linked facts, but does not transfer well to the more complex, non-document setting of dialogue memory, where the correct response is an abductive synthesis over weakly suggestive turns, no single one of which is directly relevant to the query.

3 Method

Figure 2 overviews the ADAR pipeline: notation (§3.1), Abductive Distance Score (§3.2), restricted-oracle trace synthesis (§3.3), filtering (§3.4), and student training and inference (§3.5).

3.1 Problem Setup

A dialogue-memory instance is a tuple (q, H, y^*), where q is a user query, $H = \{t_1, \ldots, t_n\}$ is the user's prior dialogue history segmented into turns, and y^* is the gold response (open-form text or an MCQ letter). At training time the benchmark also provides a designated gold evidence subset $e^* \subseteq H$. We denote by A_x the set of content anchors of a text x, extracted with spaCy (Honnibal et al., 2020) via named-entity recognition and part-of-speech tagging (we keep named entities, content-noun lemmas, and content adjectives, lower-cased and stop-word filtered). We write $\mathrm{idf}(w)$ for the smoothed inverse-document-frequency weight of anchor w over the per-benchmark corpus of queries and turns, and $\mathrm{emb}(\cdot)$ for the unit-normalised Qwen3-Embedding-0.6B (Zhang et al., 2025) encoder. Throughout, $E \subseteq H$ denotes a generic evidence set whose two natural instantiations are $\{e^*\}$ (the gold) and P_r (the cumulative pool retrieved by round r during the online search).

3.2 Abductive Distance Score

ADS combines two complementary distance axes: a lexical axis $\delta_\ell \in [0, 1]$ and a contrastive semantic axis $\delta_\sigma \in [0, 2]$. A small value along an axis means the evidence set E covers the query along that axis.

Lexical distance. The lexical axis measures the IDF-weighted fraction of query anchors absent from the evidence set E:

$$\delta_\ell(q, E) = \frac{\sum_{w \in A_q \setminus A_E} \mathrm{idf}(w)}{\sum_{w \in A_q} \mathrm{idf}(w)}, \qquad (1)$$

where A_q is the query anchor set and $A_E = \bigcup_{t \in E} A_t$ aggregates anchors over evidence turns. Rare anchors carry more weight, and $\delta_\ell = 0$ when every IDF-weighted query anchor appears in E.

Semantic distance. The semantic axis is contrastive. Let $s_{\mathrm{in}}(q, E) = \max_{t \in E} \cos(\mathrm{emb}(q), \mathrm{emb}(t))$ and $s_{\mathrm{out}}(q, E) =$

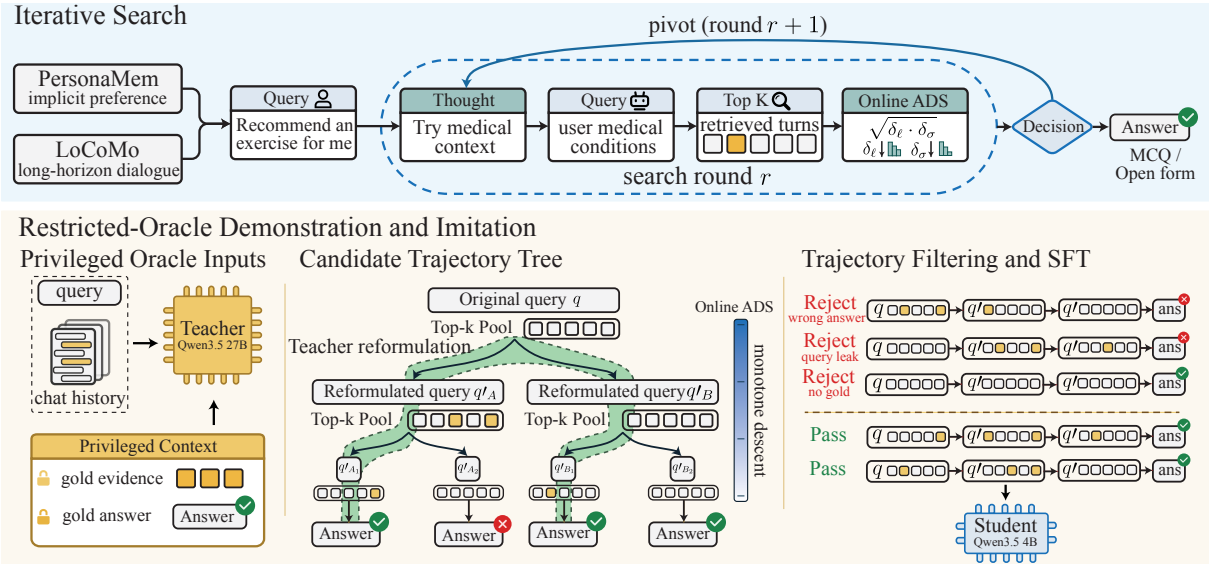

Figure 2: Overview of the ADAR pipeline. **Top:** at inference the student alternates Phase A (THOUGHT/QUERY) and Phase B (retrieve, read online-ADS metrics, commit/pivot), answering in MCQ or open-form mode. **Bottom-left:** the 27B teacher sees a privileged context (gold evidence and answer) under anti-leak rules. **Bottom-centre:** per query it grows a tree of candidate trajectories with shared parent prefixes and monotonically descending online ADS. **Bottom-right:** three filters (wrong answer, query leak, no gold turn) keep the survivors, which train the 4B student.

$\max_{t' \in H \setminus E} \cos(\mathrm{emb}(q), \mathrm{emb}(t'))$ be the best in- and out-of-evidence cosines. We define

$$\delta_\sigma(q, E) \;=\; 1 - \big(s_{\mathrm{in}}(q, E) - s_{\mathrm{out}}(q, E)\big). \quad (2)$$

When the best in-evidence turn beats the best out-of-evidence turn the distance is below one. When a distractor outside E matches the query at least as well, it exceeds one, marking E as an adversarial-retrieval region. The contrastive form is essential because absolute cosines vary across benchmarks: on HotpotQA the gold paragraph tops the ranking and δ_σ is small, while on PersonaMem-v2 distractors routinely beat the gold and δ_σ exceeds one, as illustrated in Figure 1.

Unified score. The Abductive Distance Score is the geometric mean:

$$\mathrm{ADS}(q, E) \;=\; \sqrt{\delta_\ell(q, E) \cdot \delta_\sigma(q, E)}. \quad (3)$$

The geometric mean enforces joint hardness: if either axis is small (one channel of coverage is enough), the score is small. Both axes are non-negative, with δ_ℓ in $[0, 1]$ and δ_σ in $[0, 2]$, so ADS lies in $[0, \sqrt{2}]$ and remains comparable across benchmarks and across instances.

Offline and online ADS. The two roles of ADS correspond to two choices of the evidence set. Taking $E = \{e^*\}$, the gold turn alone, gives the offline diagnostic $\mathrm{ADS}_{\mathrm{offline}}(q)$, which measures how hard a query is intrinsically, before any model is run. The semantic axis makes this concrete. When the gold turn is the closest turn to the query ($s_{\mathrm{in}} > s_{\mathrm{out}}$, so $\delta_\sigma < 1$), a naive top-1 retriever already lands on the evidence. When some distractor is at least as close ($s_{\mathrm{in}} \le s_{\mathrm{out}}$, so $\delta_\sigma \ge 1$), naive retrieval is actively pulled away from the evidence. Because $\mathrm{ADS}_{\mathrm{offline}}$ is defined per query, averaging it over a benchmark yields a single difficulty number for that benchmark, while reading it per query lets us sort the queries of a benchmark from easy to hard, a difficulty axis that does not depend on which method is evaluated. Taking $E = P_r$, the pool retrieved by round r, instead gives the online statistic $\mathrm{ADS}^{(r)}_{\mathrm{online}}(q)$, which uses no gold label and can therefore be read by the agent itself while it searches. As the pool grows, $s^{(r)}_{\mathrm{in}}$ can only rise and $s^{(r)}_{\mathrm{out}}$ can only fall, so $\mathrm{ADS}^{(r)}_{\mathrm{online}}$ decreases monotonically and tracks how much of the query the current pool already covers. It therefore acts at once as a progress monitor during search, a saturation signal for when to stop, and a post-hoc confidence signal whose plateau height reflects the query's intrinsic difficulty (§4.4).

3.3 Restricted-Oracle Trace Synthesis

The training pipeline produces an over-large pool of multi-turn agent trajectories under a restricted-oracle teacher (bottom-left and bottom-centre panels of Figure 2). Rather than generating trajectories independently, the teacher grows a single

exploration tree per query: nodes are search states, edges are search actions, and every root-to-leaf path is one candidate trajectory (detailed below). The teacher sees the gold preference and a short excerpt of the gold turn but is forbidden from leaking either into any emitted token. Only candidates passing the filters of §3.4 reach the student.

Two-phase agent loop. Each node of the tree is a search state, expanded by the same two-phase loop the student will run at inference, to a maximum depth of $R_{\max} = 5$ rounds. In Phase A of round r the teacher emits a `THOUGHT:` line that reflects on what the current pool already contains and reasons about where to search next, followed by a `QUERY:` line (one short search string). The system retrieves the top-5 turns by cosine similarity, deduplicates against the existing pool, and appends an `<observation>` block listing the new turns and a `<pool_metrics>` block reporting the current online ADS. In Phase B the teacher reads this observation and produces, in order, a `<judgment>` that assesses whether the pool now supports an answer and a `<decision>` of either `commit` or `pivot`, adding an `<answer>` when it commits. A `commit` closes the current branch at a leaf, while a `pivot` extends it into Phase A of round $r+1$. At $r = R_{\max}$ the system forces a commit. Crucially, this commit-or-pivot decision is the teacher's own judgement, formed by reading the reported online-ADS metrics and the retrieved turns, and the synthesis pipeline imposes no external numerical threshold.

Branching and parent reuse. The tree is grown by sampling, at each node, two alternative Phase-A actions (distinct THOUGHT/QUERY pairs), each spawning a child branch, so every internal node splits into two. Both children of a node reuse the parent prefix: they inherit the same dialogue history, cumulative pool P_r, and pool metrics up to the branch point, and diverge only from that round onward. Sibling trajectories thus share their common-ancestor rounds verbatim, and the candidate pool for a query is the set of all root-to-commit paths through this shared-prefix tree. The reuse both diversifies the search directions explored per query and keeps synthesis cheap, since a shared prefix is generated once and extended in several ways.

Restricted-oracle prompting. The teacher's initial user message includes a BACKGROUND block carrying the gold preference and a short gold-turn excerpt. Anti-leak rules in the system prompt forbid quoting, paraphrasing, or referencing the BACKGROUND in any emitted token. The teacher uses oracle knowledge only to *select* the search direction, not to write the answer in advance: the trajectory must hit the gold through plausible search, while the student, which never sees BACKGROUND, reproduces the same trajectory shape from imitation alone.

Student-view trace. Before any trajectory is written to disk we substitute the teacher's system prompt with a BACKGROUND-free version, replace the initial user message with the inference-time variant, and keep all assistant emissions exactly as produced. After the filters of §3.4, the SFT corpus contains 4,960 surviving traces drawn from a substantially larger candidate pool (4,666 from PersonaMem-v2 and 294 from the LoCoMo training conversation, following the LoCoMo split of Mem0 (Chhikara et al., 2025)).

Restricted-oracle SFT supplies supervision that outcome-RL cannot. The oracle is what makes trajectory-level supervision possible: knowing the gold, the teacher demonstrates not just the final answer but, at each round, which direction to point the search in, how to read the pool metrics, and when to commit. The filtering step then adds a second layer of supervision. Because we keep a trajectory only when its search actually reaches the gold turn before committing (§3.4), every surviving trace is guaranteed to trace a search path that lands on the evidence, a guarantee that can only be checked with the gold labels available at construction time. An outcome-RL agent has no such labels: it receives a single reward bit per instance and must spread it across many search decisions. On dialogue memory this signal is especially sparse, so RL tends to collapse onto trivially short trajectories or needlessly long retrievals.

3.4 Filtering

The three filters below select trajectories from the candidate pool (bottom-right of Figure 2). Only candidates passing all three reach the SFT corpus.

Anti-leak filter. A regex bank covering eight forbidden-phrase families (oracle markers, gold-turn references, target-answer references, search-for-the-answer phrasings, reverse-engineering markers, oracle-reference phrases, target-turn phrases, BACKGROUND-only vocabulary) is checked against every assistant message, and a se-

		PersonaMem-v2 (MCQ)						LoCoMo (OF)				
Method	Type	Neutral	A-stereo	Therapy	Health	Stereo	**Overall**	S-hop	M-hop	O-dom	Temp	**Overall**
Native RAG	Training-free	27.91	22.22	32.11	23.68	45.79	29.32 $\pm$1.38	20.59	0.78	14.46	45.25	29.91 $\pm$0.88
IRCoT	Training-free	28.49	32.16	23.31	39.47	42.06	32.37 $\pm$3.51	23.11	2.33	25.30	43.85	30.60 $\pm$1.50
Search-o1	Training-free	33.14	20.47	19.77	35.09	25.23	26.71 $\pm$2.05	21.01	4.28	31.33	37.85	27.67 $\pm$0.94
Search-R1	RL-tuned	14.57	9.15	29.49	14.14	18.48	16.64 $\pm$0.66	13.03	3.50	26.51	32.82	22.95 $\pm$0.42
DeepResearcher	RL-tuned	25.00	22.22	20.23	21.05	17.76	21.63 $\pm$0.68	26.47	5.84	27.71	46.09	33.31 $\pm$1.26
ADAR	SFT	**45.35**	**45.03**	**39.31**	**56.14**	**54.21**	**47.31** $\pm$2.42	**31.93**	**9.34**	25.30	**48.04**	**35.94** $\pm$0.57

Table 1: Main results on PersonaMem-v2 and LoCoMo. All methods use a Qwen3.5-4B backbone, and accuracies are means over three seeds with standard deviations across seeds (§4.1).

mantic detector additionally rejects traces where the `QUERY` field shares two or more multi-character tokens with the gold-turn excerpt that are absent from the user's original query.

Outcome filter. On MCQ benchmarks the teacher's final answer is re-evaluated through the same stateless letter judge that the student faces at inference, against the same shuffled options, and traces whose judge letter disagrees with the gold letter are discarded. On open-form benchmarks the answer is instead scored by the same LLM judge used at evaluation, and traces it judges incorrect are likewise discarded.

Evidence-presence filter. The trace is discarded if the cumulative pool P_{r^*} at the commit round contains no annotated gold evidence turn. Without this filter, the teacher could occasionally guess the right answer from priors before its search ever touches the relevant evidence, teaching the student to hit the right option without ever grounding it in what the user revealed. We do not filter on trajectory length, commit-round position, or pool-metrics dynamics, letting those distributions emerge naturally.

3.5 Inference

At inference (top panel of Figure 2) ADAR, the full-fine-tuned Qwen3.5-4B student, runs the same two-phase loop on the user's real history, and a `commit` ends the trajectory while the `<answer>` is scored by a stateless judge that has no visibility into the search trajectory, a separation we call the *confirm-reveal* protocol. Training hyperparameters and the benchmark-native judges are given in §4.1 and Appendix B.

4 Experiments

In this section, we conduct comprehensive experiments to answer the following research questions: **RQ1:** How does ADAR compare to other methods on dialogue-memory benchmarks? **RQ2:** Does offline ADS coherently characterize difficulty across and within benchmarks? **RQ3:** Does online ADS behave as an interpretable search signal at inference time? **RQ4:** Which components of ADAR are responsible for the gains?

4.1 Setup

Datasets and splits. On PersonaMem-v2 we use the benchmark's provided train and test splits, synthesising traces on the training set and evaluating on the test set, and we restrict evaluation to the memory-recall query types, excluding two types that test PII redaction and forgetting rather than retrieval (Appendix A). LoCoMo uses the same conversation-level train/test split as Mem0 (Chhikara et al., 2025), synthesising on the training conversation and evaluating on the held-out test conversations with zero overlap. LongMemEval is held out from training entirely, and we drop its knowledge-abstention queries, which test calibration rather than memory retrieval.

Training. ADAR and the two RL baselines all start from the same Qwen3.5-4B initialisation, update all parameters, and train on the same data, so the comparison isolates the training objective rather than data or capacity. Optimisation hyperparameters are given in Appendix B.

Protocol. PersonaMem-v2 uses four-choice MCQ accuracy under its native letter-judge. LoCoMo and LongMemEval use open-form accuracy judged by a frozen Qwen3.5-122B under a YES/NO confirm-reveal prompt. For agentic methods, the search loop runs options-blind on PersonaMem-v2: the four MCQ options never enter the agent's prompt during search, and they appear only at the final answering step, where a separate stateless model maps the agent's free-form answer to one of the four letters.

Baselines. Five 4B-class baselines span training-free and RL-tuned agentic search. **Native RAG** retrieves the top-ten turns by query cosine and answers directly. **IRCoT** (Trivedi et al., 2023) interleaves chain-of-thought with retrieval. **Search-o1** (Li et al., 2025) runs an agentic search-and-reason loop with document-grounded answering. **Search-R1** (Jin et al., 2025) and **DeepResearcher** (Zheng et al., 2025) are outcome-RL search agents. We retrain each on the same training data and from the same Qwen3.5-4B initialisation as ADAR, so the comparison isolates the training paradigm. All methods share the Qwen3.5-4B backbone, Qwen3-Embedding-0.6B retriever, and confirm-reveal judging.

Reporting. We report mean accuracy and standard deviation across three runs (different decoding seeds at temperature 0.7) on the evaluation sets described above. The overall-column deviation is measured directly across three seeds rather than propagated from per-category deviations.

4.2 Main Results

To answer **RQ1**, we compare our method with baselines. Table 1 reports per-category and overall accuracy on PersonaMem-v2 and LoCoMo, the two benchmarks that span the implicit-end and lookup-end of the trajectory described in §2. ADAR obtains the highest overall accuracy on both, with overall **47.31**% on PersonaMem-v2 (the next-best 4B method, IRCoT, sits at 32.37%) and **35.94**% on LoCoMo (DeepResearcher 33.31%). On LongMemEval ADAR reaches **48.40**% overall, again beating every 4B-class baseline. The per-category table is deferred to Appendix D.

Two patterns emerge. First, the gap between ADAR and the strongest 4B baseline is largest on the most abductive benchmark (+15% on PersonaMem-v2 vs +2.6% on LoCoMo), consistent with PersonaMem-v2's higher offline ADS (§4.3). Second, the two RL search agents underperform IRCoT, the strongest training-free agentic baseline, on every benchmark. Search-R1 in particular collapses to 16.6% on PersonaMem-v2 despite being trained on the same dialogue data as ADAR, direct evidence for the SFT-versus-RL argument of §3.3.

4.3 Offline ADS as a Difficulty Diagnostic

Cross-benchmark. Figure 3 situates seven benchmarks on the two-dimensional plane spanned by δ_σ (horizontal) and δ_ℓ (vertical). The four factual benchmarks (HotpotQA, 2WikiMultiHopQA, LongMemEval, MuSiQue) cluster at $\delta_\sigma < 1$, the implicit-preference ones (PrefEval, PersonaMem-v2) cross the $\delta_\sigma = 1$ tie line, and LoCoMo straddles the boundary. Mean offline ADS on PersonaMem-v2 is around 0.25, four times higher than MuSiQue and an order of magnitude higher than LongMemEval, consistent with the explicit-to-implicit spectrum of §2.

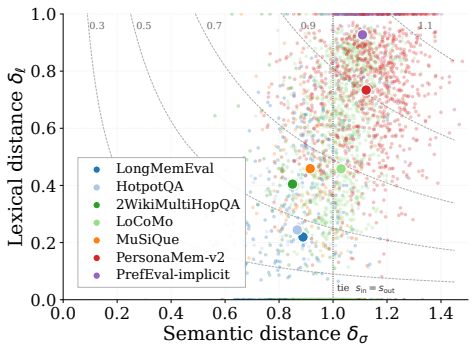

Figure 3: Cross-benchmark difficulty plane. Each scatter point is one query of one benchmark, centroids are large circles, and grey dashed curves are iso-ADS contours. Lookup-style benchmarks cluster at low δ_σ, while abductive ones cross the tie line at $\delta_\sigma = 1$.

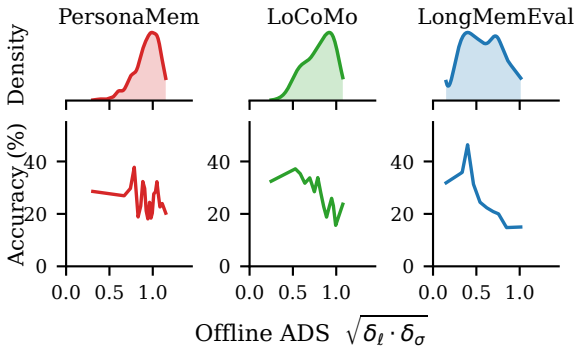

Figure 4: Within-benchmark joint plots (PersonaMem-v2 is MCQ, whereas LoCoMo and LongMemEval are open-form). Upper subplot of each panel: the distribution of offline ADS over the benchmark's queries. Lower subplot: baseline-mean accuracy across twenty equal-mass offline-ADS bins.

Within-benchmark. Figure 4 shows offline-ADS distributions and per-bin accuracy stratified by offline ADS. On LongMemEval and LoCoMo, accuracy decreases monotonically with offline ADS for every method, training-free or RL-trained, confirming that ADS captures intrinsic difficulty rather than a method-specific artefact. On PersonaMem-v2 the within-benchmark ADS distribution is concentrated in a narrow high range, so per-bin baseline accuracy is flat, and the lift pro-

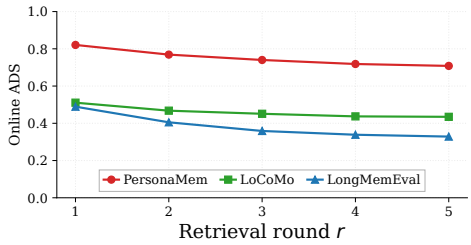

Figure 5: Online ADS descent across retrieval rounds for ADAR, averaged over the full evaluation set.

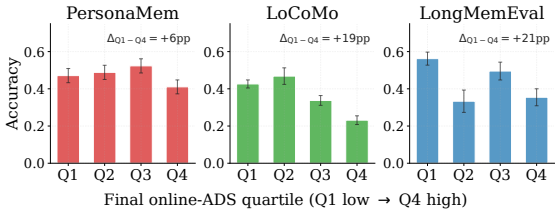

Figure 6: Calibration of accuracy against final commit-time online ADS, with bars showing equal-mass quartiles of the final ADS.

duced by ADAR is a uniform shift of the entire ADS-conditioned accuracy curve, not a selective gain on easy or hard quartiles.

4.4 Online ADS as a Search Signal

Descent. Figure 5 plots mean online ADS across retrieval rounds. The descent is strictly monotone on all three benchmarks, matching the theoretical guarantee of §3.2. Its slope is steepest on LongMemEval (-0.157 over five rounds), shallowest on LoCoMo, and intermediate on PersonaMem-v2, mirroring the cross-benchmark difficulty ranking of Figure 3: an abductive benchmark plateaus higher because cumulative retrieval struggles to push the pool past the adversarial-embedding tail.

Coupling and calibration. Stratifying PersonaMem-v2 queries by offline-ADS quartile yields a clean plateau spread: easy queries plateau at online ADS 0.630, hard queries at 0.715, the empirical form of the limit identity in §3.2. Binning queries by their final commit-time online ADS reveals a calibrated post-commit confidence signal on the two open-form benchmarks (Figure 6): on LongMemEval, committing with final ADS in $[0, 0.1)$ gives roughly 50% accuracy versus roughly 20% for $[0.7, 1.0)$, a spread of about 30 points, and LoCoMo shows about 15. On PersonaMem-v2 the spread is narrow because the final-commit ADS distribution is itself compressed, the same "no easy questions" phenomenon as Figure 4 viewed at commit time.

Row	PM	LoCoMo	LME
ADAR	**47.31**	**35.94**	**48.40**
w/o online-ADS	40.06 (−7.3)	31.76 (−4.2)	45.43 (−3.0)
w/o SFT	30.33 (−17.0)	28.67 (−7.3)	42.92 (−5.5)

Table 2: Component ablation, with parenthesised numbers giving deltas against ADAR.

4.5 Component Ablation

Table 2 ablates two pillars of the method. The **w/o online-ADS** row replaces the student-emitted commit decision with a fixed-$r = 5$ schedule (the student still emits the same Phase B blocks, but the loop ignores the `commit` until the final round). The **w/o SFT** row uses the untrained Qwen3.5-4B base model under the same inference protocol so that the only change is the absence of restricted-oracle supervision, isolating its contribution.

Both ablations degrade accuracy on every benchmark with consistent ordering ADAR $>$ w/o online-ADS $>$ w/o SFT. The largest drop is on PersonaMem-v2, where removing SFT costs 17%, and the smallest is the online-ADS ablation on LongMemEval (-3%), the most lookup-like benchmark where retrieval saturates quickly regardless of stopping policy. SFT is necessary across the spectrum, while learned commit timing contributes proportionally more on the abductive end. On average the learned agent commits around round three rather than the forced five of the ablation, and stopping once the pool has saturated keeps the answering context short and free of distractor turns that the model would otherwise have to reason over.

5 Conclusion

We introduced the Abductive Distance Score (ADS), one bounded score that doubles as an offline difficulty diagnostic on gold evidence and an online retrieval-saturation signal on the cumulative pool, together with a restricted-oracle SFT pipeline in which a gold-aware teacher generates candidate trajectories that are filtered to a focused corpus for training the dialogue search agent ADAR. Across benchmarks spanning the explicit-to-implicit spectrum, ADAR outperforms both training-free and RL search baselines. The difficulty ordering revealed by offline ADS reappears in the online-ADS descent, and online ADS at commit time tracks answer correctness. We expect the sample-and-filter paradigm to generalise where trajectory-level supervision is achievable but outcome reward is sparse.

Limitations

ADS is relative to its underlying instruments. Both axes of ADS are computed from specific components: anchors from a spaCy pipeline, an IDF table estimated on each benchmark's own corpus of queries and turns, and cosine similarities from Qwen3-Embedding-0.6B. The score is therefore a relative diagnostic rather than an absolute scale. The numerical values, and in borderline cases the cross-benchmark ordering, can move under a different anchor extractor, embedding model, or IDF corpus. ADS should be read as a comparison made under a fixed instrument, not as a benchmark-independent constant.

The training recipe requires gold-evidence annotations. Our restricted-oracle teacher needs the gold preference and a gold-turn excerpt to steer its search, and the evidence-presence filter needs annotated gold turns to reject lucky guesses. The pipeline therefore applies only to benchmarks that provide such annotations at training time and cannot be run on raw conversation logs without them. This is the price of trajectory-level supervision, and it distinguishes our setting from outcome-RL, which needs only a final reward.

Scope of the comparison. All systems share a single 4B backbone and a 27B teacher, and we do not study how the conclusions change with student or teacher scale. The two RL baselines were designed and tuned for factual multi-hop QA over shared corpora, and we adapt and retrain them on dialogue memory, which may understate their best-case behaviour.

Ethical Considerations

We used AI assistant with language editing and improving the clarity of the manuscript. The authors reviewed and verified all AI-assisted content and are responsible for the final paper.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations (ICLR)*.

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*.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, Huajun Chen, and Ningyu Zhang. 2025. LightMem: Lightweight and efficient memory-augmented generation. *arXiv preprint arXiv:2510.18866*.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777. Association for Computational Linguistics.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 6609–6625. International Committee on Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

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: Human Language Technologies (NAACL-HLT)*, pages 7036–7050. Association for Computational Linguistics.

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

Bowen Jiang, Yuan Yuan, Maohao Shen, Zhuoqun Hao, Zhangchen Xu, Zichen Chen, Ziyi Liu, Anvesh Rao Vijjini, Jiashu He, Hanchao Yu, Radha Poovendran, Gregory Wornell, Lyle Ungar, Dan Roth, Sihao Chen, and Camillo Jose Taylor. 2025b. PersonaMem-v2: Towards personalized intelligence via learning implicit user personas and agentic memory. *arXiv preprint arXiv:2512.06688*.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering

research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025. Search-o1: Agentic search-enhanced large reasoning models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5420–5438. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2024. MemGPT: Towards LLMs as operating systems. In *Conference on Language Modeling (COLM)*.

Bo Peng, Zhiheng Wang, Heyang Gong, and Chaochao Lu. 2025. IP-Dialog: Evaluating implicit personalization in dialogue systems with synthetic data. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. Association for Computational Linguistics.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711. Association for Computational Linguistics.

Qwen Team. 2026. Qwen3.5. https://huggingface.co/collections/Qwen/qwen35.

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025. R1-Searcher: Incentivizing the search capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2503.05592*.

Hao Sun, Zile Qiao, Jiayan Guo, Xuanbo Fan, Yingyan Hou, Yong Jiang, Pengjun Xie, Yan Zhang, Fei Huang, and Jingren Zhou. 2025. ZeroSearch: Incentivize the search capability of LLMs without searching. *arXiv preprint arXiv:2505.04588*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037. Association for Computational Linguistics.

Md Nayem Uddin, Kumar Shubham, Eduardo Blanco, Chitta Baral, and Gengyu Wang. 2026. From recall to forgetting: Benchmarking long-term memory for personalized agents. *arXiv preprint arXiv:2604.20006*.

Shu Wang, Edwin Yu, Oscar Love, Tom Zhang, Tom Wong, Steve Scargall, and Charles Fan. 2026. MemMachine: A ground-truth-preserving memory system for personalized AI agents. *arXiv preprint arXiv:2604.04853*.

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 *The Thirteenth 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. *arXiv preprint arXiv:2603.26680*.

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)*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2369–2380. Association for Computational Linguistics.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. 2025. PrefEval: Do LLMs recognize your preferences? evaluating personalized preference following in LLMs. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. 2025. DeepResearcher: Scaling deep research via reinforcement learning in real-world environments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 414–431. Association for Computational Linguistics.

Wanjun 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*, volume 38, pages 19724–19731.

A Benchmark Details

We evaluate on three benchmarks chosen as representative points of the explicit-to-implicit spectrum.

PersonaMem-v2 is a four-choice MCQ benchmark for implicit preference inference in task-grounded dialogue. Its queries span seven preference types, and we evaluate the memory-recall slice of five behavioural types (neutral, anti-stereotypical, therapy-background, health-and-medical, stereotypical) and drop two types that do not test memory recall. `sensitive_info` (about 10% of items) places a PII string in the query itself and asks the model to redact it, so correctness requires masking rather than retrieval. `ask_to_forget` (about 21%) asks the model to honour a prior "forget X" instruction, so correctness requires suppressing a stated fact, again not retrieval. The five retained types form the abductive behavioural-retrieval core.

LongMemEval tests long-horizon factual recall over multi-session chat. We evaluate on the 470 memory-recall questions, excluding the 30 knowledge-abstention questions whose correct answer is to decline to answer, since these test calibration rather than memory retrieval. We report four families: single-session, multi-session, temporal, and knowledge-update. LongMemEval is used only for evaluation, never for training.

LoCoMo is an open-form question-answering benchmark over long-term, multi-session conversations. We follow the conversation-level split of Mem0 (Chhikara et al., 2025), with disjoint conversations for trace synthesis, validation, and evaluation and zero overlap between synthesis and test. We report single-hop, multi-hop, open-domain, and temporal questions.

B Training and Inference Details

Student training. We full-fine-tune the Qwen3.5-4B base model (Qwen Team, 2026) on the 4,960-trace corpus with effective batch size eight, learning rate 1×10^{-5} under a cosine schedule, two epochs, and a 12,288-token context. Loss is causal-LM cross-entropy with role masking, so gradients flow only on assistant tokens. The two RL baselines start from the same initialisation, also update all parameters, and use the same training data (Appendix C).

Confirm-reveal judging. At inference the committed `<answer>` is scored by a stateless judge that has no visibility into the search trajectory: a four-choice MCQ letter judge for PersonaMem-v2, and a frozen Qwen3.5-122B YES/NO judge for LongMemEval and LoCoMo, both benchmark-native. The judge sees only $(q, \text{options}, \text{candidate})$ or $(q, \text{gold}, \text{candidate})$, which prevents the search prompt from contaminating answer extraction. Judge prompts are in Appendix E.

Compute and infrastructure. All experiments run on a single node with 4× NVIDIA A100 80GB GPUs. The models involved are the Qwen3.5-4B student and 4B-class baselines, the Qwen3.5-27B synthesis teacher, the Qwen3-Embedding-0.6B retriever, and the Qwen3.5-122B judge. The full pipeline, covering trace synthesis, student and RL-baseline training, and evaluation across all three benchmarks and three seeds, takes roughly 15 days of wall-clock time in total.

C Baseline Configurations

All five baselines use the Qwen3.5-4B backbone and the Qwen3-Embedding-0.6B retriever, and are judged under the same confirm-reveal protocol as ADAR.

Training-free. **Native RAG** retrieves the top-10 turns by query cosine and answers in a single call (greedy decoding, 300 max new tokens). **IRCoT** (Trivedi et al., 2023) interleaves chain-of-thought with retrieval for up to five iterations, retrieving the top-3 turns per step and terminating when a reasoning step emits "so the answer is". **Search-o1** (Li et al., 2025) runs an agentic search-and-reason loop for up to five search rounds, top-5 turns per search, with an in-documents reasoning step that condenses retrieved turns before answering.

Outcome-RL (retrained). **Search-R1** (Jin et al., 2025) and **DeepResearcher** (Zheng et al., 2025) are retrained from the Qwen3.5-4B base with full-parameter GRPO on the same trace corpus as our method, each with its own outcome reward. Both use identical training hyperparameters for a fair comparison: 500 GRPO steps, learning rate 1×10^{-6}, KL coefficient $\beta=0.001$, and 4 rollouts per

step. At inference, both retrieve up to 5 rounds with top-5 results per round. Both are trained with full parameters to match ADAR's training regime.

D Per-Category Results with Standard Deviation

The three tables below report per-category mean±std. Means match the main-text tables, and standard deviations are over three seeds ({11, 22, 33}, temperature 0.7).

Method	Neutral	A-stereo	Therapy	Health	Stereo	**Over.**
Native RAG	27.91 ±1.01	22.22 ±1.75	32.11 ±4.96	23.68 ±8.36	45.79 ±4.24	29.32 ±1.38
IRCoT	28.49 ±4.64	32.16 ±8.65	23.31 ±5.99	39.47 ±4.27	42.06 ±5.78	32.37 ±3.51
Search-o1	33.14 ±1.01	20.47 ±4.42	19.77 ±2.75	35.09 ±5.26	25.23 ±7.35	26.71 ±2.05
Search-R1	14.57 ±3.65	9.15 ±3.65	29.49 ±1.37	14.14 ±3.04	18.48 ±3.21	16.64 ±0.66
DeepResearcher	25.00 ±1.01	22.22 ±1.19	20.23 ±4.12	21.05 ±1.52	17.76 ±0.32	21.63 ±0.68
ADAR	**45.35** ±4.64	**45.03** ±4.42	**39.31** ±6.30	**56.14** ±6.62	**54.21** ±5.78	**47.31** ±2.42

Table 3: PersonaMem-v2 per-category MCQ accuracy (%), mean±std. Categories: Neutral, Anti-stereotypical, Therapy, Health, Stereotypical.

Method	S-hop	M-hop	O-dom	Temp	**Over.**
Native RAG	20.59 ±2.59	0.78 ±0.33	14.46 ±8.33	45.25 ±2.39	29.91 ±0.88
IRCoT	23.11 ±3.76	2.33 ±2.09	25.30 ±8.33	43.85 ±1.31	30.60 ±1.50
Search-o1	21.01 ±4.80	4.28 ±0.70	31.33 ±8.00	37.85 ±2.87	27.67 ±0.94
Search-R1	13.03 ±1.72	3.50 ±0.70	26.51 ±8.33	32.82 ±1.08	22.95 ±0.42
DeepResearcher	26.47 ±5.38	5.84 ±1.39	27.71 ±2.31	46.09 ±1.31	33.31 ±1.26
ADAR	**31.93** ±1.49	**9.34** ±1.20	25.30 ±6.93	**48.04** ±0.30	**35.94** ±0.57

Table 4: LoCoMo per-category open-form accuracy (%), mean±std, judged by Qwen3.5-122B. Categories: Single-hop, Multi-hop, Open-domain, Temporal.

Method	S-Sess	M-Sess	Temp	K-Upd	**Over.**
Native RAG	58.97 ±3.46	36.94 ±1.44	20.00 ±5.99	63.16 ±4.81	44.02 ±1.48
IRCoT	66.00 ±1.15	15.15 ±1.44	17.65 ±4.96	44.29 ±8.33	37.90 ±2.59
Search-o1	49.33 ±1.15	13.13 ±0.12	11.76 ±1.37	28.57 ±6.36	27.63 ±0.64
Search-R1	22.44 ±3.06	10.81 ±1.44	12.00 ±0.17	9.21 ±6.36	14.74 ±2.25
DeepResearcher	22.44 ±10.39	10.81 ±2.50	15.20 ±0.21	32.89 ±4.81	19.44 ±3.23
ADAR	**58.67** ±4.16	**43.43** ±2.89	**33.61** ±4.96	58.57 ±7.22	**48.40** ±3.57

Table 5: LongMemEval per-category open-form accuracy (%), mean±std, judged by Qwen3.5-122B on the 470 non-abstention questions. The single-session subcategories are merged into Single-Sess (S-Sess).

E Prompts

The teacher synthesis system prompt is shown in Box 1. The student inference prompt is the same protocol with the anti-leak block removed (Box 3). The `<pool_metrics>` feedback reports the current online ADS of the cumulative pool.

Box 1 — Teacher system prompt (restricted-oracle synthesis)

```
You are demonstrating a personalized-memory search agent.
      The agent searches a user's chat history to answer a
      query.

PROTOCOL -- strict two-phase round structure:

Phase A (when asked for THOUGHT/QUERY): output EXACTLY two
       lines:
THOUGHT: <one sentence on which angle to try next,
      grounded in what you can observe>
QUERY: <a single short search string>

Phase B (after the system returns observation +
      pool_metrics): output the judgment block, using ONLY
      these tags:
<judgment> 1-3 sentences over the online-ADS reading +
      content evidence </judgment>
<decision>commit|pivot</decision>
<answer> open-form answer, grounded in retrieved turns </
      answer>  (commit only)

pool_metrics reports the current online ADS of the
      cumulative pool: a lower value means the pool now
      covers the query well (retrieval has saturated); a
      higher value means relevant evidence is still missing
      .

You have up to 5 rounds. If you reach round 5 without
      committing, commit on round 5 with the strongest
      available evidence.

==== ANTI-LEAK RULES (load-bearing) ====
You may be given background context about the user's
      preference. Your job is to demonstrate how an agent
      -- which has NO such context at inference time --
      would naturally reach the supporting evidence.

RULE 1  Search queries must be plausible given ONLY the
      user's query. Use synonyms, related concepts, natural
       pivots. Do NOT use keywords from the background
      unless they would naturally arise from the query.
  BAD : query "easy ways to cut down on sugar"; background
       mentions "Hebrew festival recipes"; QUERY: "Hebrew
      festival low sugar".
  GOOD: QUERY: "low-sugar baking habits".
RULE 2  Forbidden words/phrases (case-insensitive); any
      discards the trace: "gold", "oracle", "hint", "told",
       "provided", "given", "target answer", "ground truth
      ", "the context says", "target turn", "answer-bearing
       turn", and any phrasing implying out-of-band
      knowledge of the preference.
RULE 3  THOUGHT must read as an in-the-moment hypothesis,
      not knowledge.
  BAD : "The target turn likely contains 'visiting
      researchers'..."
  GOOD: "Earlier searches found general topics; let me try
       the user's work context."
RULE 4  Cite retrieved turns by their turn# (seen in <
      observation>). Never reference a turn not in the
      current pool.
RULE 5  The QUERY must not contain a multi-word phrase
      that appears neither in the user's query nor in what
      you have already retrieved.

If tempted to use background info to shortcut the search,
      DON'T. Run the search as an authentic agent would,
      and let the retrieved evidence drive your judgment.
```

Box 2 — Teacher initial user message (BACKGROUND block)

```
User query: {query}

BACKGROUND (only for your synthesis judgment -- NEVER
      paraphrase, quote, or reference in output):
- general preference being tested: {preference}
- a turn supporting this preference exists in the chat
      history (you must find it through natural search; do
      NOT copy phrases from the next line)
  excerpt-for-judgment-only: '{gold_snippet_head}'

Use the BACKGROUND ONLY to verify when you have found
      supporting evidence -- never to construct your
      queries or thoughts.

Phase A -- output THOUGHT and QUERY for round 1.
```

Box 3 — Student inference prompt (system + round messages)

```
[system]
You are a personalized-memory search agent. You search a
      user's chat history to answer a query.

PROTOCOL -- strict two-phase round structure:
Phase A: output EXACTLY two lines:
THOUGHT: <one sentence on which angle to try next>
QUERY: <a single short search string>
Phase B (after observation + pool_metrics): output, using
      ONLY:
<judgment> 1-3 sentences over the online-ADS reading +
      evidence </judgment>
<decision>commit|pivot</decision>
<answer> open-form answer grounded in retrieved turns </
      answer> (commit only)
pool_metrics reports the current online ADS of the
      cumulative pool (lower = pool covers the query well;
      higher = evidence still missing).
You have up to 5 rounds; at round 5 commit with the best
      evidence.

[user, round 1]
User query: {query}

Output THOUGHT and QUERY for round 1.

[user, after each search]
<observation>
[turn#{idx} sim={cos}] {turn text} ...
</observation>
<pool_metrics>online_ADS={value}</pool_metrics>

Judge the pool and decide commit/pivot.

[user, on pivot]
Output THOUGHT and QUERY for the next round.
```

Box 4 — MCQ letter judge (PersonaMem-v2)

```
[system]
You answer a multiple-choice question based on a candidate
       answer. Pick the option (a/b/c/d) that best matches
      the candidate answer. Reply with exactly one letter.

[user]
Question: {query}

Options:
(a) {option a}
(b) {option b}
(c) {option c}
(d) {option d}

Candidate answer: {answer_text}

Which option (a/b/c/d) best matches the candidate answer?
      Reply with one letter only.
```

Box 5 — YES/NO confirm-reveal judge (LongMemEval, LoCoMo)

```
You are an evaluation judge for a personalized-response
      benchmark.

Given the user's query, the hidden ground-truth user
      preference, and the model's response, decide if the
      response REFLECTS / HONORS the preference. The
      response need not mention the preference verbatim --
      it just needs to be consistent with what someone
      holding that preference would receive as a helpful
      answer.

Output exactly one token: YES or NO.

User's query: {user_query}
Hidden ground-truth user preference: {gt_preference}
Model's response: {model_response}

Decision (YES or NO):
```