DISTILLING ACROSS TASKS INTO A GENERALIZABLE MEMORY POLICY FOR PERSONALIZED AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Personalized agents must use a user's history across changing tasks and domains. Existing memory systems typically rely on hand-designed routines or task-specific training, leaving open how to learn a shared, transferable memory policy. We introduce MEMORYOPD (Memory On-Policy Distillation), which trains a small student across tasks using the token distributions of specialist teachers operating within a memory harness. On MemoryCD, this harness combines episodic evidence with a semantic user profile; distillation transfers the teachers' memory-conditioned behavior on the student's own generations, providing dense token-level supervision without a downstream task reward. Our analysis characterizes a geometric teacher consensus at a common context and gives sufficient conditions for a memory-conditioned policy to transfer across domains. We evaluate generalization through two complementary settings with separately trained policies. A policy trained jointly on MemoryCD tasks is reused without parameter updates on eight held-out product domains. Separately, a policy trained on LongMemEval is transferred to LoCoMo without further training. In domain, MEMORYOPD configurations outperform the seven comparison methods on all eight personalization metrics for both student families, including a 27% reduction in Qwen rating MAE (0.490 to 0.357). The transferred MemoryCD policies lead the eight-domain averages on every metric; the LongMemEval policy reaches 0.695 held-out accuracy (best baseline 0.629) and transfers successfully to LoCoMo. These results support learning memory behavior that balances multiple tasks and remains useful beyond its training distribution.

1 INTRODUCTION

Personalized language agents increasingly accompany users across sessions, tasks, and domains (Salemi et al., 2024; Zhang et al., 2025a; 2026). Their usefulness depends on using long histories of interactions within finite context windows (Liu et al., 2024). A memory harness retrieves episodic evidence and consolidates semantic information about the user; the agent must then identify which information matters and combine it into a task decision. We call this learned behavior a *memory policy*. Its operations may be expressed through memory-conditioned outputs rather than explicit retrieval or update commands. Agent-memory surveys emphasize this connection between accumulated experience and useful behavior over time (Huang et al., 2026; Zhang et al., 2025b).

Many memory systems retain redundant histories or apply fixed procedures for using them (Xia et al., 2026; Fang et al., 2026). Common choices include recency windows (Tan et al., 2025), retrieval (Maharana et al., 2024), gist summaries (Lee et al., 2024), and evolving personas (Zhang et al., 2025a). These procedures can incorporate new information, but their memory-use rules are not optimized across tasks. Another line learns memory actions from downstream rewards (Wang et al., 2025; Yan et al., 2025; Zhou et al., 2025; Yu et al., 2026). Such training can improve task performance, while requiring a reward and leaving transfer beyond the training distribution to be established. We study a complementary approach: distilling task specialists into a shared memory policy and evaluating its behavior on held-out domains and a new benchmark.

The central challenge is to balance task-specific decisions with reusable memory behavior. Rating an item, ranking candidates, and writing a review require different outputs, yet all can draw on the same episodic evidence and semantic user preferences (Zhang et al., 2026). A policy that learns to use this

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

information across tasks may also remain useful when the product category or conversational setting changes.

We address this challenge with MEMORYOPD (Memory On-Policy Distillation). For each task, a specialist teacher operates within the memory harness. A shared student then matches the teachers' next-token distributions on its *own* generations (Agarwal et al., 2024; Gu et al., 2024). The targets reflect the teachers' use of episodic and semantic memory, allowing supervision of memory-conditioned behavior without requiring a separate label for every internal operation. This provides dense distributional feedback across generated positions, rather than a scalar downstream reward. A warm-start supplies initial task capabilities, and cross-task distillation aims to balance them in one policy.

Our analysis addresses two questions. First, what target does reverse-KL matching to several teachers define at a common context? The answer is their normalized geometric mean (Theorem 1); its relation to a useful shared policy depends on the teachers' aggregate bias (Corollary 1). Second, under what conditions can a memory-conditioned policy transfer? A representation-level bound relates target risk to source imitation error, feature-distribution shift, and conditional-risk shift (Theorem 2). We use these results to state conditions, and evaluate task balance and transfer through downstream experiments.

We test two complementary forms of generalization (Figure 1). Within MemoryCD, one policy per student–teacher configuration serves all four tasks and transfers without parameter updates to eight held-out product domains. Separately, a policy trained on the LongMemEval training split is evaluated on its held-out questions and transferred to LoCoMo. The results include a 27% reduction in Qwen in-domain rating MAE and LongMemEval accuracy of 0.695, compared with 0.629 for the best baseline. Our contributions are:

- Methodologically, we introduce MEMORYOPD, which distills the memory-conditioned behavior of per-task specialists into a shared student using on-policy top-K reverse KL. Dense token-level supervision supports learning across tasks without a downstream reward (Section 3, Algorithm 1).
- Theoretically, we characterize the idealized common-context reverse-KL target and give a conditional transfer bound for the full memory-conditioned policy. The analysis distinguishes teacher bias and top-K approximation error from the empirically observed task performance (Section 4).
- Empirically, across two student families, six teacher sizes, four personalization tasks, eight held-out product domains, and two conversational benchmarks, we evaluate joint task performance and transfer. The component ablation compares OPD with SFT and examines warm-start initialization and episodic+semantic memory (Tables 1 to 4, Figure 4).

2 PROBLEM SETUP

User, memory, and the memory policy. A user history $\mathcal{M}_u = \{(i, m_i, r_{u,i}, t_{u,i}, x_{u,i})\}$ records items, metadata, ratings, timestamps, and review or message text (Hou et al., 2024; Zhang et al., 2026). On MemoryCD, a harness supplies query-relevant episodic evidence $E_u(x)$ and a semantic user profile S_u, following the episodic–semantic organization of persona-based agents (Zhang et al., 2025a). We write

$$E_u(x) = R_{\text{epi}}(x, \mathcal{M}_u), \quad S_u = R_{\text{sem}}(\mathcal{M}_u), \quad \hat{y} \sim \pi_\theta(\cdot \mid c), \quad c = (x, E_u(x), S_u). \tag{1}$$

The *memory policy* selects and combines information from these inputs to produce task outputs; its operations need not appear as explicit commands. Teachers use the same harness organization, so their token distributions supervise this memory-conditioned behavior. Harness-provided retrieval and profile construction are specified separately. For conversational benchmarks, c denotes the query and the benchmark's memory interface.

Tasks and metrics. We use MemoryCD (Zhang et al., 2026), built on real Amazon Reviews 2023 users (Hou et al., 2024), with four user-facing tasks that probe different memory abilities: rating prediction (MAE and RMSE, lower better), review summarization (ROUGE-L, BLEU-1) and review generation (ROUGE-L, BLEU-4), and item ranking over a candidate pool (NDCG@1 and NDCG@3), all higher better except the rating errors (Lin, 2004; Wang et al., 2013). Main-text summaries lead with rating MAE, ranking NDCG@3, and ROUGE-L; the full eight-metric tables are

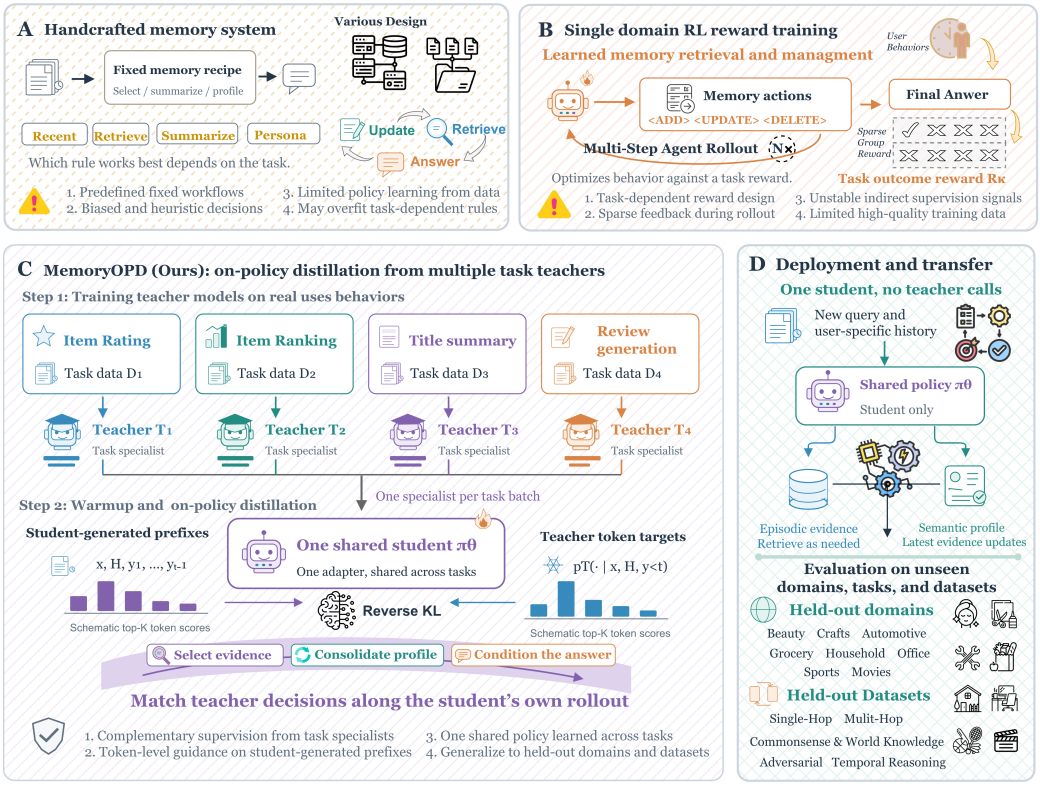

Figure 1: **Distilling across tasks into a transferable memory policy.** **(A)** Hand-crafted memory systems use fixed procedures for recency, retrieval, gist, or persona construction. **(B)** Task-reward training supervises memory behavior through downstream outcomes. **(C)** MEMORYOPD distills task specialists operating within a memory harness into a shared student using reverse-KL targets on the student's own generations. **(D)** We evaluate two separate transfer settings: the MemoryCD-trained policy on eight held-out product domains, and a separately LME-trained policy on LoCoMo. Each source-trained policy is reused without parameter updates along its corresponding transfer path.

reported per setting. For long-horizon conversational memory we use LongMemEval (LME) with its GPT-4o judge over six question types (Wu et al., 2025), and LoCoMo with word-F1, token-F1, and an LLM judge over its four answerable question categories (Maharana et al., 2024). All numbers are leakage-guarded: teacher hints and gold labels are never in the student's prompt, and evaluation uses a held-out test set that never steers training.

3 METHOD: MEMORYOPD

MEMORYOPD distills how task specialists use memory into one shared student (Figure 1). Episodic evidence and semantic profiles support different outputs across tasks; teacher distributions provide dense supervision for using that information. Matching these distributions on student-generated contexts trains the policy on the prefixes it actually produces, beyond supervision on a fixed gold sequence.

Distillation objective. Given memory context c, sample and detach $y \sim q_{\bar{\theta}}(\cdot \mid c)$ from the current student. At position t, let $p_S = \pi_\theta(\cdot \mid c, y_{<t})$ and $p_T = \pi_{\phi_k}(\cdot \mid c, y_{<t})$. Renormalizing both on $B_t = \text{top-}K(p_S)$ gives

$$\ell_t = \sum_{v \in B_t} \tilde{p}_S(v) \log \frac{\tilde{p}_S(v)}{\tilde{p}_T(v)} = \mathrm{KL}(\tilde{p}_S \| \tilde{p}_T), \tag{2}$$

$$\mathcal{L}(\theta;\bar{\theta}) = \mathbb{E}_{y\sim q_{\bar{\theta}}}\left[\frac{1}{|\mathcal{T}(y)|}\sum_{t\in\mathcal{T}(y)} c_t \ell_t\right], \tag{3}$$

where $\mathcal{T}(y)$ indexes generated positions and c_t is a unit-mean teacher-confidence weight (Appendix D). Reverse KL matches the teacher's distribution shape within the student's candidate set (Gu et al., 2024; Ko et al., 2024); gradients pass through student probabilities, not the sampled trajectory (Agarwal et al., 2024).

Sharing across tasks. Task-specific teachers supervise one parameter set in a round-robin schedule. A warm-start supplies initial capabilities, while OPD aims to balance their subsequent improvement through dense distributional feedback. Memory operations are learned through the teachers' harness-conditioned behavior, without requiring explicit operation labels. Algorithm 1 in Appendix D gives the training loop; at deployment, the student uses the memory harness without teacher calls.

OPD training. Each update pairs a source task with its specialist teacher and generates a fresh student rollout. The teacher scores the student's prefixes, providing dense feedback on the decisions encountered under the current policy. After updating the shared parameters, we regenerate rollouts so that supervision follows the evolving student. Algorithm 1 summarizes this loop; benchmark-specific schedules and optimization settings are given in Section 5 and Appendix D.

4 Analysis of Shared Memory Policies and Transfer

We analyze the full memory-conditioned policy of (1). The common-context reverse-KL identity below characterizes an idealized multi-teacher target; the transfer bound in Appendix B gives sufficient conditions for low target risk.

4.1 Cross-task distillation admits a reverse-KL consensus

Fix a shared memory state h (a context) and drop it from the notation. Let teacher k induce next-token distribution P_k, the student Q, and task weights $w_k > 0$ with $\sum_k w_k = 1$. In the *reverse* (mode-seeking) direction of Section 3, the multi-task objective *at that state* is $J_{\mathrm{rev}}(Q) = \sum_k w_k \,\mathrm{KL}(Q\|P_k)$.

Theorem 1 (Reverse-KL consensus decomposition)**.** *Assume the teachers have a nonempty common support. Let $G(a) = \frac{1}{Z}\prod_k P_k(a)^{w_k}$ be the normalized weighted* geometric *mean of the teachers, with $Z = \sum_a \prod_k P_k(a)^{w_k}$ and $D_{\mathrm{rev}} = -\log Z$. Then for every student Q supported on $\bigcap_k \mathrm{supp}(P_k)$,*

$$\underbrace{\sum_k w_k \,\mathrm{KL}(Q\|P_k)}_{\textit{multi-task distillation}} = \underbrace{D_{\mathrm{rev}}}_{\textit{teacher disagreement}} + \underbrace{\mathrm{KL}(Q\|G)}_{\textit{student gap}}, \tag{4}$$

so the unique minimizer over all admissible distributions is $Q^\star = G$ and the minimum value is the irreducible disagreement $D_{\mathrm{rev}} \geq 0$ (with $D_{\mathrm{rev}} = 0$ iff all P_k coincide).

At a common state, reverse-KL matching combines teachers into their normalized geometric mean G. Teacher disagreement determines both the constant D_{rev} and this shared target, while the student gap measures how closely a policy approximates it.

Relation to task-conditioned training. Algorithm 1 visits one task k and its teacher at a time. Ignoring top-K truncation, confidence weights, and optimization regularizers, its simplified distillation objective is

$$\sum_k w_k \,\mathbb{E}_{h\sim d_{q,k}}\left[\mathrm{KL}\big(Q(\cdot \mid h)\|P_k(\cdot \mid h)\big)\right], \tag{5}$$

where $d_{q,k}$ is the state distribution visited by the student rollout policy on task k. The effective teacher weights at a state are proportional to $w_k d_{q,k}(h)$, with additional token weighting in the implemented objective. Task-identifying prompts can yield states supervised by a single specialist; overlapping task occupancies yield multi-teacher targets of the form in Theorem 1. For a stage-dependent schedule, the task weights are $w_k^{(s)}$ for stage s; generation-focused refinement changes these weights, and its configured token weighting further changes the effective objective.

From consensus to memory behavior. Writing each teacher as a reference policy $\pi^\star$ times a task-dependent bias, the geometric target equals $\pi^\star$ when the aggregate bias is action-independent. Corollary 1 quantifies departures through the bias variation η. This identifies a condition under which combining specialists yields a useful shared policy. For top-K matching, Lemma 1 further decomposes the full-policy gap into conditional distribution shape, support mass, and a tail residual.

4.2 Conditions for transfer

Let Q denote the full stochastic memory-conditioned policy, and let ϵ bound its mean next-token KL to a reference policy $\pi^\star$ on its own source-domain trajectories. For outputs of at most H tokens and normalized bounded task loss, Theorem 2 gives

$$R_T(Q) \le R_S(\pi^\star) + \sqrt{H\epsilon/2} + L_Q W_1(\phi_\# S, \phi_\# T) + \Delta_{\mathrm{cond}}. \tag{6}$$

Here ϕ is a policy representation, W_1 measures feature-distribution shift, and Δ_{cond} is the conditional-risk shift. The risk covers the complete generated behavior, including the final answer. The bound links transferable memory behavior to accurate source imitation and stable representations and conditional risks across domains.

Scope. The analysis characterizes a common-context target and sufficient conditions for transfer. Applying it to a trained student requires accounting for task occupancies, model capacity, aggregate teacher bias, and top-K residuals. The transfer bound assumes matching token distributions and state occupancies for one execution policy; top-p training and greedy evaluation require their respective imitation-error conditions (Appendix B). Our experiments assess task balance and held-out transfer; internal bias cancellation and the bound's terms remain unmeasured.

5 Experiments

We first examine whether one policy balances the source tasks, then test product-domain transfer and its component contributions. A separately trained conversational policy tests benchmark transfer; cross-family distillation probes the limits of the distillation interface.

Setup. Students are Qwen3.5-4B-Instruct (Qwen Team, 2026) and Gemma-4-E4B (Team et al., 2026), with same-family specialist teachers at $\{4, 9, 27\}$B and $\{\text{E4B}, 26, 31\}$B, respectively. We compare four hand-designed memory systems: RecentMem (Tan et al., 2025), LoCoMo (BM25) (Maharana et al., 2024), ReadAgent (Lee et al., 2024), and PersonaAgent (Zhang et al., 2025a). Learned baselines are Mem1 (Zhou et al., 2025), Memory-R1 (Yan et al., 2025), and MemAgent (Yu et al., 2026); Table 5 situates the memory approaches. MemoryCD evaluates 300 held-out examples per task, LME uses 151 held-out questions from a $349/151$ split, and LoCoMo uses 1540 answerable questions. Table 7 specifies their isolation and transfer protocols. MemoryCD and LME policies are trained separately, with one student per backbone, teacher-size, and source-benchmark configuration. Fixed training endpoints are 150 steps for LME and 270 for the reported MemoryCD recipe, including 120 overgate updates after the default 150-step multi-task stage (Appendix D).

5.1 In-domain results

Table 1 tests the first requirement: a shared policy should perform well across all four tasks, rather than improve one at the expense of the others.

The best baseline varies by task, whereas MemoryOPD maintains strong performance across the eight metrics. Each Qwen configuration exceeds the comparison methods on all eight, including a 27% reduction in rating MAE with same-size teachers (0.490 to 0.357). Gemma's mid-size configuration also leads all eight; its other configurations lead seven, with RMSE slightly above Memory-R1. For Qwen, same-size teachers favor rating and ranking, while mid-size teachers yield the strongest summarization and generation scores. Gemma favors mid-size teachers on rating and ranking and same-size teachers on the two text tasks. Larger teachers provide no consistent gain. Figure 2 clarifies how the balance develops: rating and summarization improve, generation rises before fluctuating, and ranking stays near the level established by the warm-start. The joint result therefore combines preserved initial capabilities with improvements during distillation.

Table 1: **In-domain distillation (MemoryCD A-S1), full metrics.** Methods (columns) are grouped into hand-crafted policies, RL-trained memory agents, and MEMORYOPD; tasks/metrics are rows. The three MEMORYOPD columns are a same-, mid-, and large-size teacher *from the student's own family*; the concrete teacher sizes are listed in each student band (Qwen: 4B/9B/27B; Gemma: E4B/26B/31B). Rating is lower-better, all others higher-better. Each MEMORYOPD column is one invocation at a fixed leakage-free endpoint. Per row, **best** (bold) and second (underlined) are marked.

	Hand-crafted memory				RL memory agents			MEMORYOPD (ours)		
	Recent Mem	LoCoMo	Read Agent	Persona Agent	Mem1	Memory -R1	Mem Agent	same-size	mid-size	large-size
Qwen3.5-4B student								**4B**	**9B**	**27B**
Rating: MAE	0.4900	0.5067	0.5233	0.5333	0.6900	0.5167	0.5467	**0.3567**	0.4033	0.3667
Rating: RMSE	0.8103	0.8446	0.8775	0.8679	1.1930	0.8869	0.9747	**0.6952**	0.7371	0.7257
Ranking: NDCG@1	0.2433	0.2600	0.3467	0.3800	0.4567	0.1967	0.2600	**0.5767**	0.5600	0.5700
Ranking: NDCG@3	0.4879	0.5018	0.5727	0.5942	0.6397	0.4475	0.4921	**0.7044**	0.6961	0.7015
Summ.: ROUGE-L	0.1423	0.1342	0.1346	0.1120	0.1216	0.1412	0.1263	0.1875	**0.1928**	0.1843
Summ.: BLEU-1	0.1209	0.1140	0.1124	0.0948	0.1002	0.1171	0.1056	0.1645	**0.1678**	0.1578
Gen.: ROUGE-L	0.1317	0.1329	0.1303	0.1217	0.1211	0.1277	0.1227	0.1416	**0.1445**	0.1419
Gen.: BLEU-4	0.0193	0.0192	0.0191	0.0173	0.0173	0.0180	0.0177	0.0231	**0.0237**	0.0232
Gemma-4-E4B student								**E4B**	**26B**	**31B**
Rating: MAE	0.5533	0.5400	0.5600	0.5833	0.5667	0.6233	0.6367	0.4733	**0.4667**	0.4700
Rating: RMSE	1.0376	1.0231	1.0520	1.0247	0.9869	0.9183	0.9574	0.9238	**0.9165**	0.9220
Ranking: NDCG@1	0.5633	0.5333	0.5867	0.6000	0.4800	0.2700	0.3600	0.6733	**0.6767**	0.6667
Ranking: NDCG@3	0.6982	0.6820	0.7090	0.7224	0.6502	0.4443	0.5635	0.7685	**0.7701**	0.7639
Summ.: ROUGE-L	0.1740	0.1714	0.1452	0.1422	0.1192	0.1151	0.1028	**0.2070**	0.1967	0.2033
Summ.: BLEU-1	0.1402	0.1363	0.1200	0.1141	0.0992	0.0941	0.0850	**0.1779**	0.1665	0.1722
Gen.: ROUGE-L	0.1360	0.1330	0.1172	0.1169	0.1221	0.1261	0.1229	**0.1505**	0.1495	0.1496
Gen.: BLEU-4	0.0212	0.0206	0.0189	0.0185	0.0171	0.0188	0.0184	**0.0251**	0.0246	0.0247

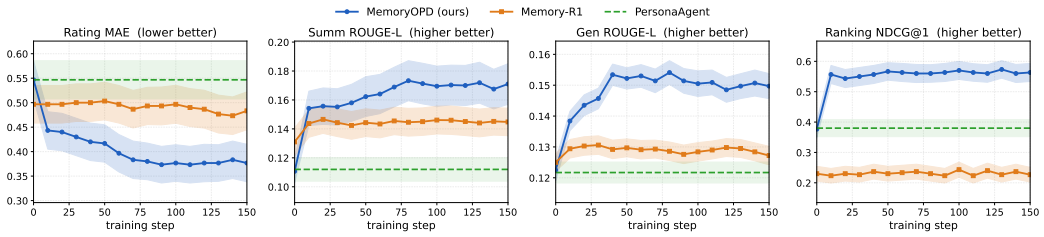

Figure 2: **Training dynamics on held-out testing** (Qwen-4B, in-domain, n=300/task; bands $\pm$1 SE). The plot covers the initial 150-step multi-task stage. Memory-R1 starts from its own Manager+Answer harness, and PersonaAgent is inference-only. MEMORYOPD improves rating and summarization, while generation rises and then fluctuates.

5.2 DOMAIN TRANSFER

Strong source-task performance leaves open whether the policy remains useful when the domain changes. We therefore apply each MemoryCD policy from Table 1 to eight unseen product domains without parameter updates; Table 2 reports the domain averages.

The transferred policies exceed the baselines on all eight mean metrics for both students. With same-size teachers, Qwen ranking reaches NDCG@3 0.613, compared with 0.467 for the best baseline, and summarization ROUGE-L reaches 0.214 versus 0.158. Gemma also improves rating MAE (0.443 versus 0.530) and generation ROUGE-L (0.149 versus 0.133). The gains span both structured decisions and text outputs.

Per-domain results. Figure 3 examines the same-size-teacher Qwen policy across the eight held-out domains. Against PersonaAgent and Memory-R1, representing hand-designed and learned memory systems, it leads on rating MAE, ranking NDCG@3, and summarization ROUGE-L in all eight domains, and on generation ROUGE-L in six of eight. The two generation exceptions are Grocery

Table 2: **Out-of-domain transfer, no retraining (MemoryCD A-S3), full metrics.** Mean over 8 held-out domains, n=300/domain; each MEMORYOPD column is the *same* in-domain adapter of Table 1, applied unchanged. Rating rows are lower-better, all others higher-better. Per row, **best** (bold) and second (underlined) are marked.

	Hand-crafted memory				RL memory agents			MEMORYOPD (ours)		
	Recent Mem	LoCoMo	Read Agent	Persona Agent	Mem1	Memory -R1	Mem Agent	same-size	mid-size	large-size
Qwen3.5-4B student								**4B**	**9B**	**27B**
Rating: MAE	0.4917	0.4958	0.5233	0.5429	0.7042	0.5221	0.5921	0.3979	0.4079	**0.3888**
Rating: RMSE	0.8538	0.8480	0.9054	0.8866	1.2020	0.9162	0.9866	**0.7344**	0.7455	0.7371
Ranking: NDCG@1	0.1954	0.1946	0.2250	0.1992	0.3267	0.1779	0.2933	**0.3996**	0.3900	0.3946
Ranking: NDCG@3	0.3891	0.3910	0.4148	0.3998	0.4665	0.3548	0.4416	**0.6126**	0.6048	0.6083
Summ.: ROUGE-L	0.1581	0.1553	0.1494	0.1419	0.1472	0.1508	0.1537	**0.2143**	0.2122	0.2127
Summ.: BLEU-1	0.1334	0.1309	0.1245	0.1180	0.1245	0.1276	0.1304	**0.1886**	0.1862	0.1876
Gen.: ROUGE-L	0.1303	0.1298	0.1278	0.1190	0.1210	0.1241	0.1228	**0.1356**	**0.1356**	0.1352
Gen.: BLEU-4	0.0220	0.0221	0.0210	0.0187	0.0195	0.0209	0.0201	**0.0250**	**0.0250**	**0.0250**
Gemma-4-E4B student								**E4B**	**26B**	**31B**
Rating: MAE	0.5433	0.5483	0.5342	0.5296	0.5979	0.5467	0.5758	**0.4425**	0.4488	0.4541
Rating: RMSE	0.9718	0.9692	0.9414	0.9185	0.9064	0.8557	0.8480	**0.8194**	0.8263	0.8370
Ranking: NDCG@1	0.3412	0.3463	0.4075	0.4042	0.3204	0.2096	0.2637	**0.4200**	0.4188	0.4117
Ranking: NDCG@3	0.5801	0.5851	0.6042	0.6168	0.4665	0.3877	0.4331	**0.6271**	0.6267	0.6226
Summ.: ROUGE-L	0.1903	0.2013	0.1561	0.1532	0.1256	0.1209	0.1190	0.2123	**0.2145**	0.2121
Summ.: BLEU-1	0.1550	0.1631	0.1281	0.1279	0.1042	0.1012	0.0993	0.1814	**0.1836**	0.1797
Gen.: ROUGE-L	0.1302	0.1327	0.1150	0.1145	0.1181	0.1248	0.1204	**0.1490**	0.1484	0.1479
Gen.: BLEU-4	0.0227	0.0234	0.0194	0.0197	0.0191	0.0210	0.0202	**0.0281**	0.0273	0.0276

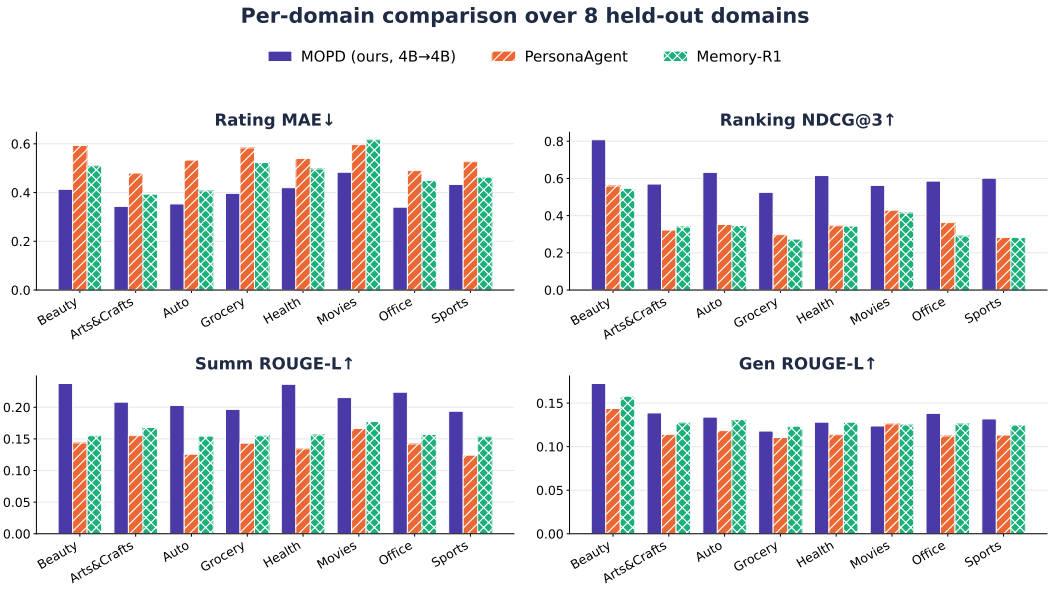

Figure 3: **Per-domain out-of-domain transfer** (8 held-out domains, n=300/domain), using one MemoryCD-trained Qwen policy without retraining. MEMORYOPD (reverse-KL, same-size 4B teacher) is compared with PersonaAgent and Memory-R1 under greedy decoding. It leads these comparators on rating MAE, ranking NDCG@3, and summarization ROUGE-L in all eight domains, and on generation ROUGE-L in six of eight.

and Movies. These results show broad transfer across product categories, with greater variation in free-text generation. Table 2 provides the aggregate comparison with all baselines, and Table 8 illustrates successful preference use in an unseen category.

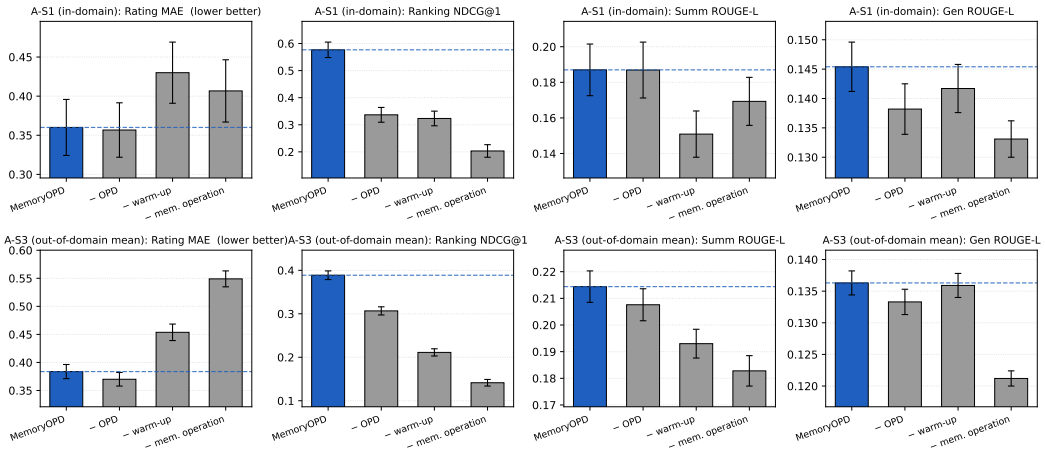

Figure 4: **Component ablation** (MemoryCD, Qwen-4B, n=300, bands ± 1 SE). −OPD replaces the distillation recipe with SFT on gold targets from the base model; −warm-up removes warm-start initialization; −mem. operation removes episodic+semantic memory from the prompt. Top row: in-domain (A-S1); bottom row: out-of-domain (A-S3; −mem. operation covers seven domains, the others eight). The dashed line marks full MEMORYOPD. The comparison illustrates the joint task balance of the complete recipe and the importance of memory inputs.

Table 3: **LongMemEval (GPT-4o judge), Qwen student.** Accuracy per question type and overall (higher-better). Evaluation details are in Appendix D. Per row, **best** (bold) and second (underlined) are marked.

	Hand-crafted memory				RL memory agents			MEMORYOPD (ours)		
	Recent Mem	LoCoMo	Read Agent	Persona Agent	Mem1	Memory -R1	Mem Agent	same-size	mid-size	large-size
Qwen3.5-4B student								**4B**	**9B**	**27B**
Temporal	0.0500	0.1500	0.3500	0.4000	0.3750	0.2000	0.3750	**0.4750**	0.3750	0.3750
Multi-session	0.1750	0.3000	0.4750	0.5250	0.5250	0.4750	0.5500	**0.7250**	**0.7250**	0.6750
Knowledge-update	**0.8333**	0.7917	0.7917	0.7500	0.7917	0.6250	0.6250	0.7917	0.7917	0.7917
Single-session user	**0.9048**	**0.9048**	0.8571	**0.9048**	**0.9048**	**0.9048**	0.8095	**0.9048**	0.8095	0.8571
Single-session asst.	0.8824	0.9412	0.9412	0.9412	**1.0000**	0.9412	0.8824	0.9412	**1.0000**	**1.0000**
Single-session pref.	0.2222	0.2222	0.4444	0.5556	0.4444	0.1111	0.3333	0.3333	**0.6667**	0.3333
Overall	0.4305	0.4901	0.5960	0.6291	0.6291	0.5166	0.5762	**0.6954**	0.6821	0.6556

Ablation. To interpret this task balance, Figure 4 compares full MEMORYOPD with −OPD (replacement by gold-target SFT), −warm-up, and −mem. operation (removal of episodic+semantic memory from the prompt). Memory removal produces the largest ranking drop (NDCG@1 falls from 0.58 to 0.20 in domain and from 0.39 to 0.14 out of domain) and reduces the other metrics. Removing warm-start lowers ranking and summarization in both scopes, while out-of-domain generation stays close to the full recipe. Compared with −OPD, full MEMORYOPD has stronger ranking and generation in both scopes, nearly identical in-domain summarization, and higher out-of-domain summarization. SFT yields slightly lower rating MAE (0.357 versus 0.360 in domain and 0.370 versus 0.384 out of domain). This supports a useful balance across tasks under the dense token-distribution supervision of OPD. The SFT run starts from the base model, so this comparison includes the full recipe's warm-start advantage (Appendix D). The out-of-domain memory-removal result covers seven domains rather than eight; its aggregate difference also reflects domain coverage.

5.3 BENCHMARK TRANSFER

Product-domain transfer changes the topic while preserving the task family. We next test conversational memory and transfer between benchmarks, using a policy trained separately on LME's six question types for 150 steps, without an additional refinement stage.

Table 4: **LoCoMo after transfer of the LME policy, Qwen student.** All 10 conversations, 1540 answerable questions; LLM-judge per category, overall judge, word-F1, token-F1 (higher-better). Per row, **best** (bold) and second (underlined) are marked.

	Hand-crafted memory				RL memory agents			MEMORYOPD (ours)		
	Recent Mem	LoCoMo	Read Agent	Persona Agent	Mem1	Memory -R1	Mem Agent	same-size	mid-size	large-size
Qwen3.5-4B student								4B	9B	27B
Multi-Hop	0.046	0.167	0.27	0.316	0.057	0.057	0.089	**0.337**	0.284	0.326
Temporal	0.028	0.259	0.402	0.399	0.034	0.019	0.084	0.411	0.399	**0.414**
Open-Domain	0.073	0.146	0.448	0.438	0.083	0.083	0.250	**0.458**	0.448	0.417
Single-Hop	0.050	0.483	0.603	0.614	0.094	0.102	0.170	0.658	0.654	**0.668**
LLM-judge (overall)	0.0442	0.3532	0.4987	0.4994	0.0760	0.0812	0.1442	0.5351	0.5201	**0.5370**
Word-F1	0.0179	0.2444	0.3275	0.3394	0.0476	0.0518	0.1081	0.3561	**0.3723**	0.3645
Token-F1	0.5384	0.6367	0.681	0.6846	0.5456	0.5347	0.4902	0.6862	**0.6982**	0.6914

On held-out LME questions (Table 3), same-size specialists yield 0.695 overall accuracy, above PersonaAgent and Mem1 (0.629). The largest gains are on temporal (0.475 vs. 0.400) and multi-session questions (0.725 vs. 0.550), while RecentMem leads knowledge-update (0.833 versus 0.792). The same-size student ties the best single-session-user score (0.905); mid- and large-size teachers tie Mem1 on assistant questions (1.000), and the mid-size configuration leads preference questions (0.667). Larger teachers again do not improve the overall result (0.695/0.682/0.656). Figure 5 provides inexpensive ROUGE-L training diagnostics; the loss generally falls, user and assistant scores rise most clearly, and temporal and preference scores show little progress (Appendix C).

Reusing the LME-trained policy on LoCoMo tests whether this behavior survives a benchmark change (Table 4). The same-size configuration exceeds all baselines on the overall judge, word-F1, token-F1, and all four categories, whereas the three RL agents score below 0.15 on the overall judge. Its overall judge score is 0.5351, compared with the best baseline of 0.4994. Across teacher sizes, the large-size configuration leads overall judge (0.5370), temporal, and single-hop scores; the mid-size configuration leads word-F1 (0.3723) and token-F1 (0.6982), while same-size teachers lead multi-hop and open-domain scores. Table 9 illustrates correct event-time answering after transfer where the two displayed comparators abstain. Parameters are unchanged, although the memory interface changes from LME persona retrieval to LoCoMo self-gist. This complements product-domain transfer with evidence from a separately trained conversational policy.

Cross-family transfer. Finally, Table 6 compares sequence-level KD, universal logit distillation, and a byte-level interface across tokenizers. For Gemma-to-Qwen transfer, byte-level matching gives the lowest rating MAE, while SeqKD leads in-domain ranking and ULD leads out-of-domain NDCG@3. For Qwen-to-Gemma transfer, SeqKD with the largest teacher leads rating and generation, while ULD and byte-level matching are stronger on summarization and NDCG@3. The preferred interface therefore varies by task and transfer direction; Appendix C details both settings.

6 Discussion and Conclusion

MEMORYOPD learns a shared memory policy by distilling task specialists operating within a memory harness. On MemoryCD, episodic evidence and semantic profiles shape the teacher behavior that the student learns from dense token-level targets on its own generations. Warm-start initialization and subsequent distillation yield a useful balance across tasks, with benefits that extend to held-out product domains. Separately, an LME-trained policy transfers to LoCoMo without parameter updates. Our analysis characterizes an idealized common-context teacher consensus and gives sufficient conditions for transfer of the full memory-conditioned policy. The experiments demonstrate cross-task balance and downstream transfer. These complementary transfer settings support learning reusable memory behavior across tasks and distributions. An open direction is to extend this behavior to explicit decisions about when to retrieve or update memory, and to distinguish the contributions of the harness and the learned policy through controlled interventions.

Reproducibility Statement

Section 3 specifies the distillation objective, and Algorithm 1 describes the training procedure. Appendix D documents the model configurations, initialization, optimization settings, training budgets, memory interfaces, evaluation metrics, and compute resources. Table 7 details the data splits and train/test separation. Appendix B provides the assumptions and proofs for the theoretical results. We will release the code as open source upon publication.

AI Use Statement

Generative AI tools primarily assisted with literature search, manuscript revision, and the analysis and organization of experimental results.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations (ICLR)*, 2024.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010.

Hongru Cai, Yongqi Li, Wenjie Wang, Fengbin Zhu, Xiaoyu Shen, Wenjie Li, and Tat-Seng Chua. Large language models empowered personalized web agents. In *Proceedings of the ACM on Web Conference 2025*, pp. 198–215, 2025.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. From persona to personalization: A survey on role-playing language agents. *Transactions on Machine Learning Research*, 2024.

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

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, et al. Lightmem: Lightweight and efficient memory-augmented generation. In *International Conference on Learning Representations (ICLR)*, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. NIPS 2014 Deep Learning Workshop.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Wei-Chieh Huang, Weizhi Zhang, Yueqing Liang, Yuanchen Bei, Yankai Chen, Tao Feng, Xinyu Pan, Zhen Tan, Yu Wang, Tianxin Wei, et al. Rethinking memory mechanisms of foundation agents in the second half. *arXiv preprint arXiv:2602.06052*, 2026.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. In *International Conference on Machine Learning (ICML)*, 2024.

Kuang-Huei Lee, Xinyun Chen, Hiroki Furuta, John Canny, and Ian Fischer. A human-inspired reading agent with gist memory of very long contexts. In *International Conference on Machine Learning*, pp. 26396–26415. PMLR, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

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

Wenhan Ma, Jianyu Wei, Liang Zhao, Hailin Zhang, Bangjun Xiao, Lei Li, Qibin Yang, Bofei Gao, Yudong Wang, Rang Li, Jinhao Dong, Zhifang Sui, and Fuli Luo. Mopd: Multi-teacher on-policy distillation for capability integration in llm post-training. *arXiv preprint arXiv:2606.30406*, 2026.

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

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Qwen Team. Qwen3.5: Towards native multimodal agents. February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: a temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*, 2025.

Ievgen Redko, Amaury Habrard, and Marc Sebban. Theoretical analysis of domain adaptation with optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pp. 737–753. Springer, 2017.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. Lamp: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7370–7392, 2024.

Haoran Tan, Zeyu Zhang, Chen Ma, Xu Chen, Quanyu Dai, and Zhenhua Dong. Membench: Towards more comprehensive evaluation on the memory of llm-based agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19336–19352, 2025.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. A theoretical analysis of ndcg type ranking measures. In *Conference on learning theory*, pp. 25–54. PMLR, 2013.

Yu Wang, Ryuichi Takanobu, Zhiqi Liang, Yuzhen Mao, Yuanzhe Hu, Julian McAuley, and Xiaojian Wu. Mem-alpha: Learning memory construction via reinforcement learning. *arXiv preprint arXiv:2509.25911*, 2025.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations*, 2025.

Peng Xia, Jianwen Chen, Hanyang Wang, Jiaqi Liu, Kaide Zeng, Yu Wang, Siwei Han, Yiyang Zhou, Xujiang Zhao, Haifeng Chen, Zeyu Zheng, Cihang Xie, and Huaxiu Yao. Skillrl: Evolving agents via recursive skill-augmented reinforcement learning. *arXiv preprint arXiv:2602.08234*, 2026.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*, 2025.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Jinhe Bi, Kristian Kersting, Jeff Z. Pan, Hinrich Schütze, Volker Tresp, and Yunpu Ma. Memory-r1: Enhancing large language model agents to manage and utilize memories via reinforcement learning. *arXiv preprint arXiv:2508.19828*, 2025.

Hongli Yu, Tinghong Chen, Jiangtao Feng, Jiangjie Chen, Weinan Dai, Qiying Yu, Ya-Qin Zhang, Wei-Ying Ma, Jingjing Liu, Mingxuan Wang, and Hao Zhou. Memagent: Reshaping long-context llm with multi-conv rl-based memory agent. In *International Conference on Learning Representations (ICLR)*, 2026.

Weizhi Zhang, Xinyang Zhang, Chenwei Zhang, Liangwei Yang, Jingbo Shang, Zhepei Wei, Henry Peng Zou, Zijie Huang, Zhengyang Wang, Yifan Gao, et al. Personaagent: When large language model agents meet personalization at test time. *arXiv preprint arXiv:2506.06254*, 2025a.

Weizhi Zhang, Xiaokai Wei, Wei-Chieh Huang, Zheng Hui, Chen Wang, Michelle Gong, and Philip S Yu. Memorycd: Benchmarking long-context user memory of llm agents for lifelong cross-domain personalization. *arXiv preprint arXiv:2603.25973*, 2026.

Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model-based agents. *ACM Transactions on Information Systems*, 43(6):1–47, 2025b.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19724–19731, 2024.

Zijian Zhou, Ao Qu, Zhaoxuan Wu, Sunghwan Kim, Alok Prakash, Daniela Rus, Jinhua Zhao, Bryan Kian Hsiang Low, and Paul Pu Liang. Mem1: Learning to synergize memory and reasoning for efficient long-horizon agents. *arXiv preprint arXiv:2506.15841*, 2025.

A Related Work

Personalized agents and hand-designed memory. Personalized agents use user histories to tailor predictions and actions (Salemi et al., 2024; Cai et al., 2025; Chen et al., 2024; Zhang et al., 2025a). The memory front-end is typically one of a recency window (Tan et al., 2025), retrieval over the raw pool (Maharana et al., 2024; Lewis et al., 2020), gist compression (Lee et al., 2024), an evolving persona (Zhang et al., 2025a), or a knowledge store (Zhong et al., 2024; Packer et al., 2023; Chhikara et al., 2025; Rasmussen et al., 2025; Xu et al., 2025). These approaches differ in how they adapt: recency and retrieval use fixed selection rules, ReadAgent constructs gist memories and selectively revisits the source, and PersonaAgent optimizes a user-specific persona prompt at test time using textual feedback. Prompt adaptation is distinct from jointly training model parameters across tasks. Our experiments compare implementations of these memory procedures with a shared, trained memory-conditioned policy.

Learned memory management. Reinforcement learning provides one route to train memory behavior. Mem-α (Wang et al., 2025) learns explicit memory operations across several task categories and evaluates distribution and length shifts. Memory-R1 (Yan et al., 2025) trains a memory manager and an answer agent, reporting zero-shot transfer from LoCoMo to MSC and LongMemEval. MEM1 (Zhou et al., 2025) jointly consolidates memory and reasoning into a compact state, with longer-horizon evaluation and zero-shot transfer to online web QA. MemAgent (Yu et al., 2026) learns to overwrite a bounded memory while processing document chunks, and evaluates length extrapolation and held-out long-context tasks. These results establish that reward-trained memory behavior can generalize beyond its source setting. MemoryOPD supplies a complementary training signal: token distributions from task specialists operating within the memory harness. It trains a shared memory-conditioned generation policy, while retrieval and profile construction remain part of the harness.

Table 5: **Where MEMORYOPD sits.** Optimization and evaluation scope of representative memory approaches. Signals refer to method-specific adaptation, excluding backbone pretraining. Prior-work entries summarize the cited methods; the fixed window and BM25 rows describe the baseline variants used here. Task coverage, held-out-domain transfer, cross-benchmark transfer, and length extrapolation are listed separately rather than treated as equivalent capabilities. For MEMORYOPD, product-domain and conversational transfer use separately trained source policies, each reused without parameter updates.

Method	Memory behavior / optimization	Adaptation signal	Task scope / reported transfer
RecentMem (Tan et al., 2025)	Fixed recency window	None	Reusable rule; source-trained transfer is not applicable.
BM25 retrieval (Maharana et al., 2024)	Fixed retrieval (LoCoMo-style baseline)	None	Reusable rule; source-trained transfer is not applicable.
ReadAgent (Lee et al., 2024)	Gist construction and selective lookup; prompting	None (prompting)	QA and summarization benchmarks; additional web-navigation evaluation.
PersonaAgent (Zhang et al., 2025a)	Per-user persona; test-time prompt optimization	Textual feedback from user history	Multiple personalization tasks; per-user adaptation.
Mem-α (Wang et al., 2025)	Explicit memory operations; RL	Task outcome and auxiliary rewards	Retrieval, classification, summarization; out-of-distribution and length evaluation.
Memory-R1 (Yan et al., 2025)	Memory manager and answer agent; RL	QA outcome reward	Multiple QA types; LoCoMo $\rightarrow$ MSC / LongMemEval, zero-shot.
MEM1 (Zhou et al., 2025)	Joint memory and reasoning state; RL	Task outcome reward	Multi-objective QA and WebShop settings; QA $\rightarrow$ online web QA, zero-shot; horizon extrapolation.
MemAgent (Yu et al., 2026)	Chunk-wise memory overwrite; multi-conversation RL	Verifiable answer reward	Long-context QA; length extrapolation and held-out long-context tasks.
MEMORYOPD	Shared memory-conditioned generation; OPD	Teacher token distributions	Four personalization tasks; held-out product domains. Separately, LongMemEval $\rightarrow$ LoCoMo.

Knowledge distillation and on-policy distillation. Distillation transfers a teacher's output distribution into a student (Hinton et al., 2015). On-policy variants use student-generated contexts with reverse KL (Gu et al., 2024), generalized distillation objectives (Agarwal et al., 2024), or streamlined losses (Ko et al., 2024). Concurrent work by Ma et al. (2026) also integrates specialist teachers into one student and shares the acronym MOPD; we use the full name MEMORYOPD to distinguish our focus on memory-conditioned behavior and transfer. Our teachers operate within a memory harness, allowing dense supervision of behavior informed by episodic evidence and semantic profiles. We analyze a common-context reverse-KL target and sufficient conditions for transfer, and test generalization with separate product-domain and conversational-benchmark transfer settings. Table 5 distinguishes optimization signals and reported evaluation settings. The student receives dense token-distribution supervision, while specialist teachers and the warm-start use task supervision.

B PROOFS AND ADDITIONAL THEORY

This appendix analyzes the complete memory-conditioned output policy in Section 4. Fixed-context results suppress $h = (c, y_{<t})$ from the notation. The training objective (3) detaches the sampled trajectory, so it has no score-function term. The common-context objective $J_{\mathrm{rev}}(Q) = \sum_k w_k \mathrm{KL}(Q \| P_k)$ is an idealization with all teachers defined at the same context. The transfer result below additionally specifies the source occupancy and the stochastic policy to which it applies.

B.1 THEOREM 1: REVERSE-KL CONSENSUS DECOMPOSITION

Proof. Write $G(a) = \frac{1}{Z}\prod_k P_k(a)^{w_k}$ with $Z = \sum_a \prod_k P_k(a)^{w_k}$, so $\log G(a) = \sum_k w_k \log P_k(a) - \log Z$. For any Q supported on $\bigcap_k \mathrm{supp}(P_k)$,

$$\begin{aligned}\sum_k w_k \,\mathrm{KL}(Q\|P_k) &= \sum_a Q(a)\Big[\log Q(a) - \sum_k w_k \log P_k(a)\Big] \\ &= \sum_a Q(a)\big[\log Q(a) - \log G(a) - \log Z\big] = \mathrm{KL}(Q\|G) - \log Z, \end{aligned} \tag{7}$$

which is (4) with $D_{\mathrm{rev}} = -\log Z$. By the weighted arithmetic-geometric mean inequality, $\prod_k P_k(a)^{w_k} \le \sum_k w_k P_k(a)$, so $Z \le \sum_a \sum_k w_k P_k(a) = 1$ and $D_{\mathrm{rev}} = -\log Z \ge 0$, with equality iff all P_k coincide. Since $\mathrm{KL}(Q\|G) \ge 0$ with equality iff $Q = G$ and D_{rev} is independent of Q, the minimizer is $Q^\star = G$ with value D_{rev}. □

Forward multi-task KL $\sum_k w_k \mathrm{KL}(P_k\|Q)$ is minimized by the arithmetic barycenter $\sum_k w_k P_k$; the reverse direction MEMORYOPD uses is minimized by the geometric mean G. This is why the consensus is G, not the arithmetic mean.

B.2 COROLLARY 1: SHARED-POLICY FACTORIZATION

Corollary 1 (Consensus is the shared policy when task biases cancel)**.** *Write each teacher on the common positive support as $P_k(a) = \pi^\star(a)e^{b_k(a)}/A_k$, so by Theorem 1 the geometric consensus is*

$$G(a) = \frac{\pi^\star(a)\, e^{\bar b(a)}}{\mathbb{E}_{a'\sim\pi^\star} e^{\bar b(a')}}, \qquad \bar b = \textstyle\sum_k w_k b_k. \tag{8}$$

If the aggregate bias is action-independent, $\bar b(a) \equiv c$, then $G = \pi^\star$ and $J_{\mathrm{rev}}(Q) - D_{\mathrm{rev}} = \mathrm{KL}(Q\|\pi^\star)$. More generally, if $\max_a \bar b(a) - \min_a \bar b(a) \le \eta$, then

$$\mathrm{KL}(Q\|\pi^\star) \;\le\; \mathrm{KL}(Q\|G) + \eta \;=\; J_{\mathrm{rev}}(Q) - D_{\mathrm{rev}} + \eta. \tag{9}$$

Proof. On the common positive support, $\prod_k P_k(a)^{w_k} = \pi^\star(a)e^{\bar b(a)}\prod_k A_k^{-w_k}$ with $\prod_k A_k^{-w_k}$ constant in a (using $\sum_k w_k = 1$); normalizing gives (8). If $\bar b \equiv c$ then $G = \pi^\star$, so $J_{\mathrm{rev}}(Q) - D_{\mathrm{rev}} = \mathrm{KL}(Q\|\pi^\star)$ by Theorem 1. In general $\log \frac{G(a)}{\pi^\star(a)} = \bar b(a) - \log \mathbb{E}_{\pi^\star} e^{\bar b} \le \max_a \bar b - \min_a \bar b \le \eta$ (since $\log \mathbb{E}_{\pi^\star} e^{\bar b} \ge \min_a \bar b$), so $\mathrm{KL}(Q\|\pi^\star) = \mathrm{KL}(Q\|G) + \mathbb{E}_Q[\log(G/\pi^\star)] \le \mathrm{KL}(Q\|G) + \eta$, which is (9). □

D_{rev} measures teacher disagreement, $\mathrm{KL}(Q\|G)$ the student's gap to their geometric target, and η the variation of their aggregate bias relative to a reference policy. Teachers can be identical yet share a bias ($D_{\mathrm{rev}} = 0, \eta > 0$), or disagree yet aggregate to $\pi^\star$ ($D_{\mathrm{rev}} > 0, \eta = 0$). Both G and D_{rev} depend on the teachers. The bias condition specifies when this target approximates a useful shared reference policy.

B.3 LEMMA 1: TOP-K SET DECOMPOSITION

Lemma 1 (What top-K reverse distillation keeps)**.** *Let $B = \text{top-}K(Q)$ be the student support the loss* (2) *uses, and renormalize every distribution on B. The on-support objective decomposes as $\sum_k w_k \mathrm{KL}(Q_B\|P_{k,B}) = D_B + \mathrm{KL}(Q_B\|G_B)$ with $G_B = G(\cdot \mid B)$, and the gap to the consensus splits as*

$$\mathrm{KL}(Q\|G) = \beta\,\mathrm{KL}(Q_B\|G_B) + \mathrm{kl}(\beta\|\gamma) + (1-\beta)\,\mathrm{KL}(Q_{B^c}\|G_{B^c}), \tag{10}$$

with $\beta = Q(B)$, $\gamma = G(B)$, and $\Omega_B = \mathrm{kl}(\beta\|\gamma) + (1-\beta)\mathrm{KL}(Q_{B^c}\|G_{B^c})$. Combining with Corollary 1 gives the per-state master bound

$$\mathrm{KL}(Q\|\pi^\star) \;\le\; \beta\,[\,J_B(Q_B) - D_B\,] + \Omega_B + \eta, \qquad J_B(Q_B) = \textstyle\sum_k w_k \mathrm{KL}(Q_B\|P_{k,B}). \tag{11}$$

Proof. Splitting $\mathrm{KL}(Q\|G)$ over the partition $\{B, B^c\}$ gives $\mathrm{kl}(\beta\|\gamma) + \beta\,\mathrm{KL}(Q_B\|G_B) + (1-\beta)\,\mathrm{KL}(Q_{B^c}\|G_{B^c})$, which is (10). Applying Theorem 1 to the renormalized teachers $\{P_{k,B}\}$ on B: their normalized geometric mean is $\propto \prod_k (P_k(a)/P_k(B))^{w_k} \propto \prod_k P_k(a)^{w_k}$, i.e. exactly G_B, so $\sum_k w_k \mathrm{KL}(Q_B\|P_{k,B}) = D_B + \mathrm{KL}(Q_B\|G_B)$. The loss (2) renormalizes both distributions on B and penalizes $\mathrm{KL}(Q_B\|G_B)$ only; it is invariant to β and to the tail Q_{B^c}. Combining (10) with (9) yields $\mathrm{KL}(Q\|\pi^\star) \le \mathrm{KL}(Q\|G) + \eta = \beta[J_B(Q_B) - D_B] + \Omega_B + \eta$, which is (11). □

Top-K residual. The loss penalizes $\mathrm{KL}(Q_B\|G_B)$, the student's conditional distribution on support B. The residual Ω_B accounts for support-mass and tail mismatch and requires separate control to bound the full-distribution gap. Concretely, fix any support mass $\beta \in (0,1)$ and let the consensus concentrate off B, $\gamma = G(B) \to 0$: then $\mathrm{kl}(\beta\|\gamma) = \beta \log \frac{\beta}{\gamma} + (1-\beta)\log\frac{1-\beta}{1-\gamma} \to \infty$ while the on-support term $\mathrm{KL}(Q_B\|G_B)$ can stay 0, so Ω_B, and with it the true gap, blows up with the on-support loss at its floor. (A vanishing support mass alone does not do this: at fixed $\gamma \in (0,1)$, $\lim_{\beta\to 0} \mathrm{kl}(\beta\|\gamma) = -\log(1-\gamma)$ is finite.) These residual terms must be controlled separately to obtain a small full-policy imitation error.

B.4 Theorem 2: imitation and cross-domain transfer

Write $\xi = (x, \mathcal{M}_u)$ for the raw input and $c = \mathcal{H}_D(\xi)$ for its memory context under the specified harness on domain D. The full stochastic policy $Q(\cdot \mid c)$ generates the trajectory, including the final answer. For a fixed measurable task-loss function $\ell \in [0,1]$, define

$$R_D(Q) = \mathbb{E}_{(\xi,y^\star)\sim D}\, \mathbb{E}_{\hat{y}\sim Q(\cdot|\mathcal{H}_D(\xi))} \ell(\hat{y}, y^\star). \tag{12}$$

Thus the risk includes answer generation as part of the complete policy. If a generated trajectory contains intermediate text, ℓ includes the fixed rule used to extract its answer.

Let both Q and the reference $\pi^\star$ emit at most H tokens, padding terminated trajectories with an absorbing end symbol. Define $d_{Q,S}$ by drawing an input from source domain S, generating a trajectory under Q, and selecting a prefix position uniformly from $1, \dots, H$. Suppose

$$\mathbb{E}_{h\sim d_{Q,S}}\, \mathrm{KL}\big(Q(\cdot \mid h)\|\pi^\star(\cdot \mid h)\big) \le \epsilon. \tag{13}$$

Where the common-context comparison of Lemma 1 applies at every such state, one possible upper bound is

$$\epsilon = \mathbb{E}_{h\sim d_{Q,S}} \left[\beta_h(J_{B,h} - D_{B,h}) + \Omega_{B,h} + \eta_h\right]. \tag{14}$$

This expression makes the dependence of ϵ on conditional fit, tail mismatch, and aggregate teacher bias explicit.

Definition 1 (Feature-factored policy)**.** A memory-conditioned policy is *feature-factored* through ϕ if $Q(\cdot \mid c) = \rho(\cdot \mid \phi(c))$. The feature may retain user preferences and task-relevant evidence. A representation that suppresses domain-specific nuisance information is a candidate for transfer, but such invariance is not assumed to follow merely from the use of a memory harness.

Theorem 2 (From policy imitation to cross-domain transfer)**.** *Assume* (13)*, the bounded loss and trajectory length specified above, and $Q = \rho \circ \phi$. On a common metric feature space, let μ_D be the distribution of $z = \phi(\mathcal{H}_D(\xi))$ for inputs from D, and assume μ_S, μ_T have finite first moments. Define the feature-conditional risk $f_D(z) = \mathbb{E}_{(\xi,y^\star)\sim D,\, \hat{y}\sim Q(\cdot|\mathcal{H}_D(\xi))}[\ell(\hat{y}, y^\star) \mid \phi(\mathcal{H}_D(\xi)) = z]$. Suppose f_S has a specified L_Q-Lipschitz extension to the union of the source and target feature supports. Then*

$$R_T(Q) \le R_S(\pi^\star) + \sqrt{\tfrac{H}{2}\epsilon} + L_Q W_1(\mu_S, \mu_T) + \Delta_{\mathrm{cond}}, \tag{15}$$

where $\Delta_{\mathrm{cond}} = \mathbb{E}_{z\sim\mu_T}|f_T(z) - f_S(z)|$ is the feature-conditional risk shift. In the main text, $\phi_\# D$ abbreviates this feature distribution with the harness included.

Proof. Step 1 (sequence divergence). Let $\mathbb{P}_{S,Q}$ and $\mathbb{P}_{S,\pi^\star}$ be joint distributions with the same source input and reference-label law, followed by generation under Q or $\pi^\star$ in the same source harness. The autoregressive KL chain rule and the uniform-prefix definition of $d_{Q,S}$ give

$$\mathrm{KL}(\mathbb{P}_{S,Q}\|\mathbb{P}_{S,\pi^\star}) = H\, \mathbb{E}_{h\sim d_{Q,S}} \mathrm{KL}(Q(\cdot \mid h)\|\pi^\star(\cdot \mid h)) \le H\epsilon.$$

Step 2 (source risk). Since the complete trajectory loss lies in $[0,1]$, its expectation difference is at most total variation. Pinsker's inequality therefore gives $R_S(Q) - R_S(\pi^\star) \le \sqrt{H\epsilon/2}$. *Step 3*

(representation transfer). By conditional expectation, $R_D(Q) = \mathbb{E}_{\mu_D} f_D$. Adding and subtracting $\mathbb{E}_{\mu_T} f_S$ yields

$$R_T(Q) - R_S(Q) = (\mathbb{E}_{\mu_T} f_S - \mathbb{E}_{\mu_S} f_S) + \mathbb{E}_{\mu_T}(f_T - f_S) \le L_Q W_1(\mu_S, \mu_T) + \Delta_{\text{cond}},$$

using Kantorovich–Rubinstein duality and the assumed extension of f_S. Combining the source-risk and transfer inequalities proves the result. □

Execution policies. The representation term follows optimal-transport domain adaptation (Ben-David et al., 2010; Redko et al., 2017) and depends on the feature metric, L_Q, and conditional-risk shift. Source and target memory interfaces enter through μ_D and f_D. To apply the theorem to top-p sampling or greedy decoding, Q and $d_{Q,S}$ must denote that execution policy and its induced occupancy, with ϵ evaluated accordingly. Relating different execution policies instead requires additional distribution-comparison conditions.

C Full result tables

This appendix gives the LongMemEval training diagnostics behind Section 5.3 and the cross-family transfer table.

LME diagnostics. Table 3 reports final LLM-judge accuracy, with the largest gains on temporal and multi-session questions. Figure 5 instead uses ROUGE-L as an inexpensive diagnostic on held-out questions, alongside training loss. The loss generally declines from about 0.86 to 0.52, with late fluctuations. User and assistant ROUGE-L rise most clearly, multi-session and knowledge-update show smaller net gains, temporal ends slightly below its initial score, and preference remains near zero. Final judge scores use the fixed training endpoint in Table 3.

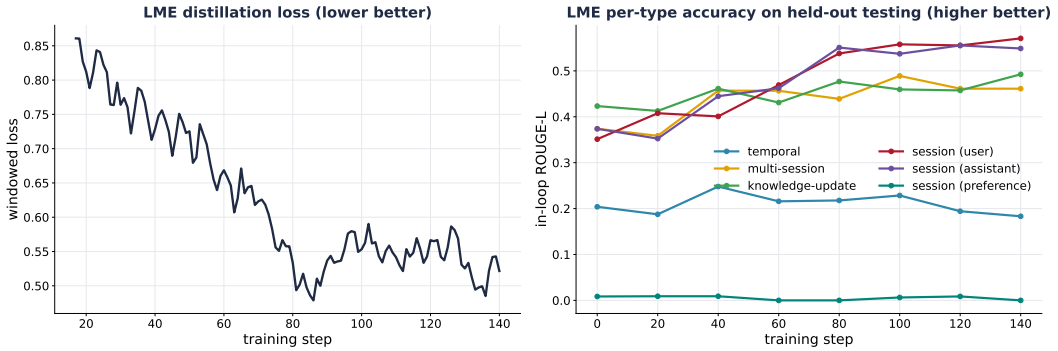

Figure 5: **LongMemEval training diagnostics.** *Left:* the distillation loss generally falls. *Right:* per-type ROUGE-L provides an inexpensive intermediate evaluation, distinct from the final GPT-4o judge in Table 3. ROUGE-L is the metric shown on the right panel's ordinate. Several types improve; temporal has no final net increase, and the nine-question preference slice stays low. The available observations extend through step 140 of the 150-step schedule.

Cross-family transfer. Table 6 compares three interfaces in both directions and in both domain settings: **SeqKD** uses teacher-generated text for SFT; **ULD** matches sorted top-K probability vectors on student rollouts; the **byte-level interface** matches distributions over a shared 256-symbol byte vocabulary. For Gemma-to-Qwen transfer, byte-level matching with the 26B teacher gives the lowest rating MAE (0.3500 in domain and 0.3752 out of domain). SeqKD with the 26B teacher leads in-domain NDCG@3 (0.7382), while ULD with the E4B teacher leads out of domain (0.6579). SeqKD also supplies the strongest generation scores in both scopes; summarization favors SeqKD with the 31B teacher in domain and byte-level matching with the E4B teacher out of domain. For Qwen-to-Gemma transfer, SeqKD with the 27B teacher leads rating MAE (0.4050 and 0.4027) and generation ROUGE-L (0.1446 and 0.1426), as well as generation BLEU-4. ULD leads in-domain summarization, while byte-level matching leads out-of-domain summarization and in-domain NDCG@3; ULD leads out-of-domain NDCG@3. Teacher-size changes are modest for ULD and

byte-level matching but less consistent for SeqKD. These results show task-dependent tradeoffs among the three cross-family interfaces.

D TEACHERS, HYPERPARAMETERS, AND PROTOCOL

Backbones. Students are Qwen3.5-4B-Instruct (Qwen Team, 2026) and Gemma-4-E4B (Team et al., 2026); the Instruct variants are used because base variants cannot reliably format the rating and ranking outputs. Qwen generation uses `enable_thinking=False` to avoid chain-of-thought consuming the answer budget. The Gemma student uses a rank-initialized adapter and text-decoder-only LoRA targets.

Parameterization and initialization. Experiments use a single LoRA adapter (Hu et al., 2022) on a frozen student backbone (rank 128, $\alpha = 256$ for Qwen; rank 64, $\alpha = 128$ for Gemma). Teachers have frozen backbones with per-task specialist adapters. For the MemoryCD warm-start comparison, TIES merges the four same-backbone task-specialist adapters to initialize the shared student; −warm-up omits this initialization. The −OPD run replaces the distillation recipe with SFT on gold targets from the base model. Its comparison with full MEMORYOPD therefore measures the reported recipes, including their initialization choices.

Confidence weighting. In (3), the teacher-confidence weight is

$$c_t = \frac{\max_{v \in B_t} \tilde{p}_{T,t}(v)}{\bar{c}}, \qquad \bar{c} = \frac{1}{|\mathcal{T}(y)|} \sum_{t \in \mathcal{T}(y)} \max_{v \in B_t} \tilde{p}_{T,t}(v).$$

The normalization gives unit mean within each generated trajectory. At update n, $\bar{\theta} = \theta_n$ denotes the student used for rollout sampling; the sampled tokens remain fixed during the gradient update.

Optimization and training budget. Each update uses a batched student rollout (batch 2, temperature 1.0, top-p 0.95) and the reverse top-K KL of (2), with $K = 16$ and the teacher-confidence weighting of (3) (`sel-token`). The learning rate is 3×10^{-6} with a cosine schedule and 0.1 warmup. Tasks are visited round-robin: four tasks for MemoryCD and six question types for LME. The default multi-task budget is 150 steps; LME uses this budget without an additional refinement stage. MemoryCD's additional processing is described below. Per-task KL normalization is off for MemoryCD and on for LME. The base-model anchor is disabled for MemoryCD; the LME configuration uses coefficient 0.3. The trainer supports a 0.3 additional weight on explicitly identified pre-answer spans (`lambda-rec`). An EMA parameter regularizer with coefficient 0.02 is used during multi-task training. Endpoints are fixed, with no early stopping or test-set checkpoint selection.

Algorithm 1 MEMORYOPD: shared-policy on-policy distillation

Require: source tasks $\mathcal{K}$; harness-conditioned teachers $\{\pi_{\phi_k}\}$; initialized student π_θ; default multi-task budget $N = 150$; top-K; experiment configuration

for $n = 1$ to N **do**
  choose task $k \in \mathcal{K}$ round-robin and draw a source input
  form memory context c using the task's memory harness
  set $\bar{\theta} \leftarrow \theta$; generate $y \sim q_{\bar{\theta}}(\cdot \mid c)$; detach y
  score each generated prefix with teacher π_{ϕ_k} and student π_θ
  form the student top-K conditional loss ℓ_t using (2)
  compute $\mathcal{L}$ using (3) and update θ
end for
if the MemoryCD refinement configuration is enabled **then**
  reset optimizer state and restart learning-rate warmup
  run 120 additional generation-focused updates using the configured task schedule and loss (Appendix D)
end if
return shared policy π_θ for source tasks and transfer evaluation

Additional MemoryCD configuration. The reported MemoryCD recipe includes a generation-focused refinement after the 150-step multi-task stage, with 120 additional overgate updates, an

Table 6: **Cross-family transfer (MemoryCD), full metrics.** A source-family teacher transfers to the *opposite*-family student by **SeqKD** (text), **ULD** (token), and a **byte-level** interface. Rating MAE lower better; all others higher. Within each block and column, **best** and second-best are shaded.

A-S1 (in-domain): Qwen to Gemma-E4B

Method (teacher)	Rating	Ranking		Summ.	Generation	
	MAE	N@1	N@3	R-L	R-L	B-4
SeqKD (4B)	0.4333	0.4967	0.6604	0.1587	0.1348	0.0122
SeqKD (9B)	0.6000	0.4800	0.6233	0.1389	0.1359	0.0119
SeqKD (27B)	**0.4050**	0.5467	0.6820	0.1544	**0.1446**	**0.0129**
ULD (4B)	0.5333	0.6700	0.7647	0.1866	0.1350	0.0105
ULD (9B)	0.5367	**0.6733**	0.7680	**0.1873**	0.1355	0.0107
ULD (27B)	0.5367	0.6700	0.7647	0.1827	0.1337	0.0109
byte (4B)	0.5133	**0.6733**	**0.7701**	0.1694	0.1383	0.0120
byte (9B)	0.5233	**0.6733**	**0.7701**	0.1659	0.1359	0.0117
byte (27B)	0.5167	**0.6733**	**0.7701**	0.1706	0.1356	0.0119

A-S1 (in-domain): Gemma to Qwen-4B

Method (teacher)	Rating	Ranking		Summ.	Generation	
	MAE	N@1	N@3	R-L	R-L	B-4
SeqKD (E4B)	0.4400	0.5133	0.6306	0.1209	**0.1361**	0.0120
SeqKD (26B)	0.4000	**0.6067**	**0.7382**	0.0835	0.1352	**0.0126**
SeqKD (31B)	0.5833	0.1633	0.4348	**0.1286**	0.1330	0.0123
ULD (E4B)	0.3567	0.5167	0.6608	0.1169	0.1226	0.0106
ULD (26B)	0.3567	0.5200	0.6620	0.1163	0.1235	0.0114
ULD (31B)	0.3583	0.5233	0.6616	0.1179	0.1231	0.0111
byte (E4B)	0.3533	0.5000	0.6521	0.1259	0.1275	0.0115
byte (26B)	**0.3500**	0.5033	0.6538	0.1214	0.1252	0.0111
byte (31B)	0.3567	0.4967	0.6454	0.1188	0.1213	0.0103

A-S3 (out-of-domain): Qwen to Gemma-E4B

Method (teacher)	Rating	Ranking		Summ.	Generation	
	MAE	N@1	N@3	R-L	R-L	B-4
SeqKD (4B)	0.4621	**0.4283**	0.5861	0.1688	0.1302	0.0124
SeqKD (9B)	0.6067	0.4079	0.5598	0.1415	0.1343	0.0129
SeqKD (27B)	**0.4027**	0.4021	0.5676	0.1741	**0.1426**	**0.0141**
ULD (4B)	0.5017	0.4092	**0.6211**	0.2009	0.1333	0.0128
ULD (9B)	0.5038	0.4092	0.6209	0.2016	0.1304	0.0127
ULD (27B)	0.5067	0.4063	0.6186	0.2026	0.1316	0.0131
byte (4B)	0.5154	0.4046	0.6188	**0.2031**	0.1309	0.0128
byte (9B)	0.5200	0.4058	0.6192	0.1995	0.1289	0.0122
byte (27B)	0.5167	0.4054	0.6184	0.2024	0.1313	0.0126

A-S3 (out-of-domain): Gemma to Qwen-4B

Method (teacher)	Rating	Ranking		Summ.	Generation	
	MAE	N@1	N@3	R-L	R-L	B-4
SeqKD (E4B)	0.4771	0.4296	0.5825	0.1160	0.1292	0.0117
SeqKD (26B)	0.4081	0.4250	0.5806	0.0896	**0.1324**	**0.0130**
SeqKD (31B)	0.6309	0.1683	0.3984	0.1159	0.1241	0.0109
ULD (E4B)	0.3964	0.4621	**0.6579**	0.1232	0.1129	0.0111
ULD (26B)	0.3902	0.4621	0.6574	0.1141	0.1133	0.0108
ULD (31B)	0.3890	0.4650	0.6577	0.1185	0.1096	0.0102
byte (E4B)	0.3798	0.4642	0.6172	**0.1269**	0.1147	0.0108
byte (26B)	**0.3752**	**0.4708**	0.6209	0.1242	0.1104	0.0104
byte (31B)	0.3790	0.4662	0.6189	0.1239	0.1119	0.0105

optimizer-state reset, and a fresh learning-rate warmup. During this refinement only, a configured support-overlap weight scales each position's loss:

$$\ell_t^{\text{refine}} = \text{sg}\left[\frac{|\text{top-}K(p_S) \cap \text{top-}K(p_T)|}{K}\right] \ell_t.$$

The MemoryCD refinement gate scales the entire loss at each token position by the observed support overlap. Tables 1 and 2 report the endpoint after $150 + 120 = 270$ updates. LME uses the standard 150-step multi-task stage. Figure 2 shows the first 150 steps of MemoryCD training.

Evaluation protocol and leakage control. MemoryCD evaluates 300 held-out examples per task using greedy decoding; rating is MAE, summarization and generation are ROUGE-L (Lin, 2004), ranking is NDCG@3 (Wang et al., 2013). LongMemEval uses persona retrieval and the GPT-4o judge with six per-type rubrics (Wu et al., 2025) over the 151 held-out questions of our $349/151$ split (Table 7); LoCoMo uses self-gist and its 1540 answerable questions across all 10 conversations, scored by word-F1, token-F1, and an LLM judge (Maharana et al., 2024). Teacher and student are evaluated in the same environment, and the prompt never contains a gold label, a teacher hint, or an oracle user profile. History-derived semantic profiles and episodic evidence are legitimate memory inputs. The test set never steers training: there is no early stopping and no checkpoint selection on it. Baselines ReadAgent (gist) and PersonaAgent synthesize with the student's own backbone, so all methods are self-contained and directly comparable.

Data protocol and train/test isolation. Table 7 gives the provenance, scale, and disjointness of every split. **MemoryCD** is built from Amazon Reviews 2023 users (Hou et al., 2024): the training pool is 223 users across four source domains (Home & Kitchen, Beauty & Personal Care, Books, Electronics), 44,598 instances balanced over the four tasks (≈11.1k each). Because all four tasks for a user are derived from that user's single interaction timeline, we split *by user*: the 100 evaluation users are held out entirely, so the in-domain test (A-S1) shares 0 users and 0 (user, interaction) pairs with training, and no target can leak from one task into another. Within a user, the conditioning memory is the first half of the timeline (`start_frac`=0.5) and the test anchors are later interactions (minimum timeline length 8, at most 50 anchors/user, 3 test anchors/user). Out-of-domain transfer (A-S3) uses eight product domains disjoint from the four training domains, n=300 per domain, with the MemoryCD-trained policy applied without parameter updates. Teachers (SFT) and the student (on-policy distillation) use the *same* training users; neither ever sees an evaluation user. **LongMemEval-S** ships 500 questions; because our method trains on LongMemEval-derived data, we partition the 500 *by question id* into 349 for training and 151 held out for evaluation, with no id shared. Per type the training/test counts are temporal $93/40$, multi-session $93/40$, knowledge-update $54/24$, single-session-user $49/21$, single-session-assistant $39/17$, single-session-preference $21/9$ (totals $349/151$, summing to the official 500); each question carries its own session haystack, so the split is at the question/haystack level, and all reported LongMemEval numbers are on the 151 held-out questions. **LoCoMo** is test-only: the policy trained on LongMemEval is transferred with no retraining and evaluated on the 1,540 answerable questions across all 10 conversations. No split uses a validation set or test-set checkpoint selection.

Training diagnostics. LongMemEval uses ROUGE-L as an inexpensive intermediate evaluation, avoiding repeated LLM-judge calls. Figure 5 reports the available diagnostic observations through step 140 of the 150-step schedule. Final task comparisons use the GPT-4o judge at the fixed training endpoint (Table 3).

Compute and reproducibility. Training uses a single node of eight H200 GPUs, with student and teacher on separate devices for long-sequence rollouts. A shared trainer is configured separately for each backbone, teacher size, and source benchmark. MemoryCD policies are reused on held-out product domains; separately trained LME policies are reused on LoCoMo. The latter uses the target benchmark's self-gist interface, with no parameter updates.

Table 7: **Data protocol and train/test disjointness.** Sources, scales, and the exact split each benchmark uses. Teacher SFT and student distillation draw from the *same* training pool; no benchmark uses a validation set or test-set checkpoint selection.

Split	Source	Scale	Disjoint from test
MemoryCD (Amazon Reviews 2023; split by user)			
Teacher train + student distill	4 source domains, 4 tasks	223 users; 44,598 inst. ($\approx 11.1\text{k} \times 4$)	*n/a*
Validation / dev	none (fixed-step stop)	*n/a*	*n/a*
Test, in-domain (A-S1)	held-out users, Home & Kitchen	100 users; 300/task (1,200)	0 users, 0 interactions
Test, out-of-domain (A-S3)	8 unseen product domains	300/task/domain	disjoint domains
LongMemEval-S (500 questions; split by question id)			
Teacher train + student distill	LongMemEval-S questions	349 questions (6 types)	*n/a*
Validation / dev	none (fixed-step stop)	*n/a*	*n/a*
Test (held-out)	LongMemEval-S questions	151 questions (6 types)	0 question ids
LoCoMo (public; transfer only, no training)			
Training	none (LME policy reused)	*n/a*	*n/a*
Test	10 conversations	1,540 answerable	test-only

E TRANSFER CASES

We contrast MEMORYOPD, Memory-R1, and PersonaAgent on two transfer settings using the available case records (Qwen3.5-4B). The tables separate recorded memory or retrieval excerpts, reported decisions, and our interpretation. Quotation marks denote excerpts preserved in the records; missing intermediate states are marked explicitly. The cases illustrate observed transfer outcomes. Mechanism attribution would require complete intermediate states and controlled interventions.

E.1 PREFERENCE TRANSFER

The first comparison concerns ranking five candidates in All_Beauty, a held-out product category. The record includes a single all-method comparison and a separate three-query trajectory summary for one user. Their sample-level correspondence is unavailable, so Table 8 presents them in separate panels.

Table 8: **Cross-domain ranking: recorded decisions and supporting memory excerpts.** Panel A compares the same five-candidate instance. Panel B reports a separate three-query group built from 40 cross-domain facts; its excerpts cannot be assigned to Panel A's instance. Shading highlights MEMORYOPD.

Stage	PersonaAgent	Memory-R1	MEMORYOPD
A Matched instance: which item is ranked first?			
Gold	Top-ranked item: `B095RYJJHY` (five candidates).		
Decision	✗ Gold item not first.	✗ Gold item not first.	✓ Gold item ranked first.
Assessment	Fails the recorded top-1 criterion; the predicted item is not preserved.	Fails the recorded top-1 criterion; the predicted item is not preserved.	Matches the recorded gold top item; the full ranking is not preserved.
B Supporting trajectory summary: three queries for one user			
Memory / retrieval	No method-specific memory excerpt is preserved for this group.	Retrieves preference-related evidence: "rated 5/5 'Leaves skin very soft' ".	Preference summary: "practical … favors compact travel-friendly electronics and natural, usable beauty items".
Recorded path	Intermediate states and group-level decisions are unavailable.	ADD/UPDATE memory bank → preference-related retrieval → incorrect top choices.	Cross-domain preference summary → ranking decisions → three top-1 hits.
Outcome	Not reported for this group.	✗ 0/3 top-1 hits.	✓ 3/3 top-1 hits.
Interpretation	**Evidence selection and the final preference decision**		
Observed behavior	Panel A records an incorrect top choice.	Panel B records preference-related retrieval followed by incorrect rankings.	The recorded summary expresses reusable preferences, and the group decisions are correct. This is consistent with preference use across categories.

Observed outcome. MEMORYOPD selects the gold item where both comparators fail in Panel A; Panel B adds a concrete contrast between preference-related retrieval and successful ranking. Panel B's counts summarize the displayed three-query group.

E.2 TEMPORAL ANSWERING

The second comparison applies the LME-trained policy to LoCoMo without retraining. The question asks for the time of a *planned* camping trip, requiring the answer to be associated with the correct person and event. Table 9 combines the matched answer comparison with the Memory-R1 retrieval pointer preserved in the accompanying conversation summary.

Table 9: **LME → LoCoMo: answering a question about a planned camping trip.** MEMORYOPD matches the reference month and year, whereas the two comparators abstain. Only Memory-R1 has a recorded retrieval pointer for this question; its retrieved text and the other methods' intermediate states are unavailable.

Question "When is Melanie planning on going camping?"
Reference answer June 2023.

Stage	PersonaAgent	Memory-R1	MEMORYOPD
Memory / retrieval	The case record does not preserve the persona or retrieved evidence.	Recorded pointer: `RELEVANT:99...` The corresponding memory-entry text is not preserved.	The case record does not preserve the memory state or retrieved evidence.
Recorded path	Intermediate state unavailable → abstention.	Retrieval pointer → "No info" in the trajectory summary; also marked as abstention in the matched answer table.	Intermediate state unavailable → "June 2023".
Decision	✗ Abstains on an answerable question.	✗ Abstains on an answerable question.	✓ "June 2023".
Interpretation	**Correct time recovery versus failure to answer**		
Observed behavior	Abstains on a question with a reference date.	Abstains after a recorded retrieval step.	Matches both the month and year in the reference. The record supports correct temporal answering after transfer.
What remains unobserved	Whether the relevant plan survived persona construction and was available to the reader.	Whether retrieval selected the correct event, and whether the answer stage could resolve its date.	Whether the date was retained explicitly or derived from a relative-time expression and a session timestamp.

Observed outcome. The transferred MEMORYOPD policy recovers an answerable event time that both comparators leave unanswered. Together with Table 8, it illustrates successful decisions on both a held-out product category and a held-out conversational benchmark.