Learning from Supervision with Semantic and Episodic Memory: A Reflective Approach to Agent Adaptation

Anonymous ACL submission

Abstract

Adapting large language models (LLMs) as agents for new tasks or domains remains a central challenge in NLP. Traditional approaches such as fine-tuning or parameter-efficient adaptation can be costly, inflexible, and opaque. In this work, we propose a flexible memory-augmented framework that enables LLM agents to continuously learn from both supervised signals and structured critiques without updating model parameters. Our framework distinguishes between *semantic* and *episodic memory*, and introduces two forms of reflective insights, instance-level *critiques* and generalizable *principles*, to capture and organize knowledge from labeled examples and their neighborhoods. We investigate how memory should be structured and how it can be effectively used to adapt agents to new scenarios. Across diverse tasks, our method yields up to 12.5% accuracy gains. We also introduce *suggestibility*, a new metric quantifying how readily models internalize feedback. Our findings highlight the promise of memory-driven reflective learning for building more adaptive and interpretable LLM agents.

1 Introduction

Large language models (LLMs) have demonstrated impressive generalization capabilities across a wide range of tasks. These AI agents rely on intelligence embedded in their pretrained parameters, and increasingly, on learning from task-specific signals, whether explicit (e.g., labeled supervision) or implicit (e.g., user interactions, feedback). A key challenge is enabling agents to continuously improve their performance and generalize to unseen domains or tasks by distilling knowledge from such signals and storing them in a reusable and interpretable form.

Traditional approaches to learning from new signals often involve updating model parameters through fine-tuning (Radford et al., 2018; Howard and Ruder, 2018) or adaptation mechanisms such as parameter-efficient methods (e.g., LoRA adapters) (Houlsby et al., 2019; Hu et al., 2022). While effective, these approaches incur computational cost, require retraining for every new signal or task, and often lack interpretability or controllability. Furthermore, they provide limited support for never-ending learning, where an agent must continuously adapt without retraining from scratch or storing large sets of models.

An alternative paradigm is memory-augmented learning (Weston et al., 2015; Zhong et al., 2024), where the underlying model remains frozen, and adaptation occurs through interaction with an external memory. This memory stores relevant task knowledge, examples, demonstrations, or explanations, that can be retrieved at inference time to inform the model's decisions. Among such approaches, in-context learning (ICL) (Dong et al., 2024) has emerged as a simple yet powerful mechanism, where the model is conditioned on a prompt consisting of a small number of examples (few-shot learning). However, directly incorporating supervised signals in the LLM context often relies on only few-shot input-output examples and tends to result in shallow pattern mimicking, due to a lack of deeper abstraction or conceptual understanding.

Recent work (Madaan et al., 2023; Yao et al., 2023; Shinn et al., 2023) has highlighted the capacity of LLMs to not only perform tasks but also critique them, generating feedback and identifying patterns of errors in their own outputs. Inspired by human tutoring, where feedback often includes explanations of mistakes and guidance for improvement, we explore whether such reflective insights can be distilled into reusable knowledge for future tasks. Instead of merely memorizing example responses, we hypothesize that an agent that inter-

All code is available here.

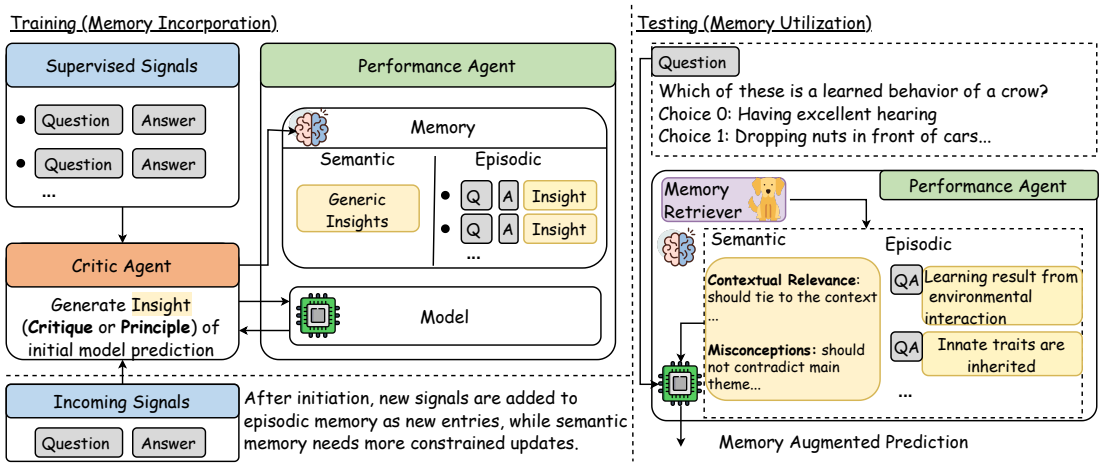

Figure 1: Agents learn from supervised signals by incorporating them into memory. At inference time, both task-level insights (semantic memory) and context-specific information (episodic memory) can support more informed decision-making.

nalizes structured feedback can develop a deeper understanding of task requirements and generalize more effectively to new examples.

In this paper, we investigate how LLM agents can effectively and continuously learn from supervised signals and deeper reasoning provided by critiques, incorporating these insights into memory. We introduce a flexible framework that models both **semantic** and **episodic memory**, and propose two forms of reflective insights: *critiques* and *principles*. These insights capture knowledge from labeled instances and their local neighborhoods, enabling more adaptive and reflective reasoning.

Grounded in this framework, our empirical investigation is guided by two central research questions: **(i) If LLM-based agents are to learn to improve continuously, then how should their memory be structured, generated, and represented? (ii) How can memory be effectively used to adapt agents to unseen scenarios?** Through extensive empirical evaluations across diverse datasets, we observe up to a **12.5% improvement in accuracy** with memory-augmented reasoning, demonstrating the effectiveness of our learning strategies. In-depth analyses reveal that while generic semantic memory is often less effective than instance-based episodic memory, combining both types yields additional benefits. We further explore how task characteristics, the components of reflective insights, and the quantity of labeled data influence performance.

Additionally, we introduce a new metric, *suggestibility*, to quantify the extent to which models internalize and adapt to critique-based insights. Our findings show that suggestibility varies by task type, with preference-oriented tasks showing greater responsiveness to critiques than fact-based tasks. We also find that the presence of rationale within critiques significantly enhances model suggestibility.

2 Learning from Supervised Signals

For large language model (LLM)-based agents, the availability of supervised signals, in the form of labeled datasets and continuously incoming feedback from users or the environment, presents a wide range of opportunities for building agents that can learn and improve continuously.

As illustrated in Figure 1, we refer to our task-solving agent as the *performance agent* (PA). The PA consists of a backbone model, which it queries to perform tasks, and a memory module, which it can read from and write to. Given a new task to which we want to adapt the agent, we begin with an initial labeled dataset:

$\mathcal{D}_{\text{train}}^{\text{init}} = \{(x_i, y_i)\}_{i=1}^{N}$, where each x_i represents a task-related question or request, and y_i denotes the corresponding correct label or answer. The performance agent processes inputs x_i from a test set $\mathcal{D}_{\text{test}}$ and produces initial predictions denoted by $\text{PA}(x_i)$.

To enhance the capabilities of the PA, we introduce a second component: the *critic agent* (CA). The CA takes as input one tuple (x_i, y_i) along with the PA's prediction $\text{PA}(x_i)$, and outputs a text cri-

tique aimed at improving the PA's performance.

3 What to Remember?

We explore how to extract useful information from supervised signals to populate the memory module. We introduce two strategies for knowledge distillation: *critiques*, which provide instance-specific feedback, and *principles*, which capture generalizable patterns to guide future predictions.

3.1 Critique

Critique is a widely used approach for improving model performance and guiding iterative refinement by identifying errors, uncovering blind spots, and providing actionable feedback for enhancement. (Shinn et al., 2023; Gou et al., 2024; Chen et al., 2024) In our setup, we employ an external, label-driven critique generation process, where the critic agent is distinct from the performance agent and uses the ground-truth answers as part of its input to generate critiques. For each question in the dataset, the performance agent first produces an initial prediction. The critique agent is then given the correct answer and asked to critique the performance agent's output. Each critique is structured into the following fields:

- **Assertion:** A reiteration of the correct answer to the question. and a judgment regarding the correctness of the performance agent's response.
- **Rationale:** An instance-specific explanation detailing why the correct answer is valid and why the performance agent's response was correct or incorrect.
- **Reflection:** A broader, generalizable insight that may be applicable to similar questions in the future.

This design addresses a key challenge observed in our empirical studies: critic agents sometimes persist in their own incorrect understanding from its underlying model when generating critiques, even after being shown the correct answer. Because the critic agent draws on the model's parametric knowledge, it can inherit pretraining biases that reinforce such confirmation. To mitigate this, we require the critic agent to explicitly restate the correct answer and make a clear assertion about the correctness of the initial prediction before offering a rationale or reflection. This explicit structure significantly reduces confirmation bias.

We further decompose critiques into two conceptual layers—*rationale* (local) and *reflection* (global)—to balance specificity and generalizability. An ideal rationale should provide a detailed explanation tailored to the specific instance, while a reflection should capture broader insights that can be applied to unseen examples in the future. This structured format resulted in noticeably higher-quality critiques.

Example Critique

Question: Which of these is a learned behavior of a crow?
Choice 0: Having excellent hearing
Choice 1: Dropping nuts in front of cars
Choice 2: Having black feathers
...
(PA chose Choice 1)
Critique:

- *Assertion:* Choice 1
- *Rationale:* Crows dropping nuts in front of cars is a learned behavior that shows their ability to adapt and mimic effective strategies, as it involves...
- *Reflection:* Learned behaviors result from experience or environmental interaction, unlike innate traits, which are inherited.

While these critiques were informative and detailed, their pointwise nature, operating on a single data point at a time, limits their ability to identify patterns across examples. To complement this, we introduce a second form of insight: *principles*.

3.2 Principle

Principles are learning artifacts designed to generalize across multiple examples. We define a principle as *"a fact, pattern, preference, or any other insight that can help answer similar questions in the future."* This intentionally flexible definition allows the critic agent to discover trends across instances. While conceptually similar to the *reflections* of critiques, principles are grounded in patterns observed across multiple examples rather than a single one.

For each data instance in the training set, we sample $K = 10$ additional instances selected based on embedding similarity using retrieval-augmented generation (RAG) (details in Appendix A.1). The $K + 1$ instances and their associated ground-truth answers are provided as input to the critic agent, which then produces a single principle that applies

to all $K + 1$ examples. This setup encourages the principles to capture patterns or insights that generalize beyond the original instance compared to the critiques.

Example Principle

Question: For user 1679, are they more likely to buy or not buy the game The Cave?
Principle: Users who have shown a preference for indie or unique gaming experiences (like "Gratuitous Space Battles" and "Hammerwatch") are more likely to buy similar games, while they tend to avoid more mainstream or well-known titles (like "Grand Theft Auto IV" and "Star Wars - Jedi Knight II Jedi Outcast").

3.3 Incorporating Insights into Memory

Next, we investigate how learned insights, whether critiques or principles, can be effectively incorporated into the agent's memory. We adopt two primary forms of agent memory: *semantic memory* and *episodic memory*, both of which are well-established in agentic learning literature (Sumers et al., 2024).

Semantic memory encodes generalizable knowledge across the entire dataset. In this work, we construct semantic memory by summarizing the contents of critiques or principles into a unified knowledge representation. This allows the performance agent to draw on abstract insights during inference in future tasks.

Episodic memory, by contrast, captures instance-specific knowledge grounded in concrete examples and the performance agent's initial behavior. For each example, we store the supervised signal (x_i, y_i) along with its corresponding critique and principle as an episodic memory entry.

4 Memory Utilization Strategies

With the addition of semantic and episodic memory, the agent gains the ability to reason using both the parametric knowledge embedded in its underlying model and the externally provided supervised signals, represented as various forms of memory. A central question that arises is how to best leverage these memories to improve decision-making during inference on new data. To this end, we explore several memory utilization strategies to study the impact of different memory representations.

4.1 Semantic Memory

Semantic memory is designed to encode generalizable insights that are broadly applicable across the entire task domain. To utilize it at inference time, we augment the performance agent's prompt with these insights in the form of additional instructions. Concretely, the `SEM_CRIT` variant incorporates summarized critiques derived from the training data, while `SEM_PRIN` uses summarized principles extracted from the same data. These prompts provide high-level guidance aimed at improving generalization on unseen examples.

4.2 Episodic Memory

While semantic memory offers concise and broadly applicable knowledge, it often fails to capture nuanced patterns or context-specific behaviors, particularly in diverse datasets. Episodic memory addresses this by enabling the agent to recall specific past instances, effectively allowing it to "revisit" similar scenarios where successes or failures occurred, along with accompanying context and critical reasoning.

The key to effective use of episodic memory lies in the retrieval of relevant examples. The agent retrieves relevant memories from similar prior cases and conditions on both the original examples and their critiques, learning to weigh and incorporate only the most pertinent critiques rather than attending to all examples equally. Following the retrieval-augmented generation (RAG) paradigm, we identify the top $K = 5$ most similar data points to a test input x_i using semantic embeddings. The corresponding memory entries, containing critical thinking artifacts are then used as additional demonstrations for the performance agent. As with semantic memory, `EP_CRIT` and `EP_PRIN` differ in the nature of their episodic content: the former emphasizes critiques on individual examples, while the latter captures localized principles distilled from a neighborhood of similar instances.

4.3 Combining Semantic and Episodic

To harness the complementary strengths of both memory types, we introduce `EP+SEM_PRIN` and `EP+SEM_CRIT`. These hybrid strategies present the performance agent with both high-level semantic instructions and context-specific episodic examples. By unifying generalizable principles with detailed situational context, these approaches aim to support more robust and adaptive reasoning during infer-

ence. These are unified by simply concatenating the semantic memory to the end of the episodic memory. There is also a variant to this approach, `EP+SEM_CRIT_LOCAL`, which creates a new semantic memory entry at prediction time. Instead of summarizing critiques over the entire training set, `EP+SEM_CRIT_LOCAL` includes a summary of only the retrieved episodic entries. This presents a trade-off between the generality of the global semantic memory, and the specificity of the local semantic memory.

5 Empirical Evaluation

5.1 Datasets

To evaluate the effectiveness of various memory-augmented learning strategies under diverse conditions, we conduct empirical studies across datasets spanning multiple domains. The tasks cover a range of settings, including fact-oriented question answering, ranking, retrieval-based QA.

MMLU (Hendrycks et al., 2021) Given a fact-based, straightforward question on one of many different domains (such as math, science, history, or logic) select the correct answer out of 4. Questions were sampled evenly across all subjects.

Multi-Condition Ranking (Pezeshkpour and Hruschka, 2025) Given a list of 5 items, sort them in order along 3 logical conditions. Converted into a 4-choice multiple-choice task.

NFCorpus (Boteva et al., 2016) Given a medical article and two medical papers, determine which paper is cited directly by the article's bibliography.

PubMed (Jin et al., 2019) Determine if a highly technical medical statement is true or false, across many different medical domains.

To mitigate potential bias from LLMs being exposed to public datasets during pretraining, we additionally evaluated our strategies on four personal preference datasets. The task was to predict whether a given item belonged to a user's history. Even if the model had encountered these datasets during training, it would be unlikely to memorize preferences associated with individual user IDs.

Steam Pref (Tamber) Video game playtime per user on the PC platform Steam. Sampled only games that were played for at least 5 hours.

Book Pref (Ziegler et al., 2005) Book ratings per user. Actual ratings were not used - the task was formatted as predicting whether a user is more or less likely to read a given title.

Anime Pref (Union) Anime ratings per user from the website MyAnimeList. Actual ratings were not used - the task was formatted as predicting whether a user is more or less likely to watch a given title.

Movie Pref (Parashar) Movie ratings per user, based on the MovieLens dataset. Sampled only movies rated 3/5 or higher.

For the preference datasets, we randomly selected three users per dataset. For each user, 250 items were sampled from their history and 250 from outside it, prioritizing favorites when possible. Users were treated independently, with no memory shared across them. For all other datasets, 500 questions were randomly sampled and evenly split into training and testing sets. Additional dataset-specific preprocessing details are provided in Appendix A.2

5.2 Experimental Setup

Most of our experiments were conducted using OpenAI's `gpt-4o-mini` (OpenAI, 2024) as the base LLM, chosen for its strong balance between cost and performance. Additionally, we ran a subset of experiments using LLaMA 4 Scout (Meta, 2025) and OpenAI's `o4-mini` (OpenAI, 2025) to examine how results vary across open-source models and compute-constrained test-time settings.

We compared our learning pipeline against two baseline setups: *zero-shot* and *few-shot*. The zero-shot baseline reflects the performance agent's output without any memory or demonstrations. The few-shot baseline includes $K = 5$ example question-answer pairs (x_i, y_i) sampled from the training set (the same as in episodic memory experiments, for consistency), where x_i is the input and y_i the corresponding answer. Note that both baselines rely solely on the original supervised signals and do not include any additional insights or memory augmentation.

5.3 Results

Table 1 compares various learning strategies against the baselines across all eight datasets. We observe substantial variance across datasets, both in baseline performance and in the effectiveness

Model and Experiment	MMLU	Multi-Condition Ranking	NFCorpus	PubMed	Steam Pref	Book Pref	Anime Pref	Movie Pref
gpt-4o-mini								
Zero-shot	88.7	56.8	85.6	**62.4**	52.8	52.0	47.9	49.9
Few-shot	88.0	**67.6**	85.6	62.0	55.8	50.9	51.1	53.2
EP_CRIT	88.0	65.2	83.6	62.0	62.7	53.8	54.4	57.7
EP_PRIN	**92.0**	66.4	85.2	61.2	**68.3**	50.7	55.1	56.4
SEM_CRIT	87.3	58.4	**87.2**	59.6	60.1	45.5	48.8	58.7
SEM_PRIN	87.3	58.4	85.6	59.2	62.0	51.5	46.3	57.1
EP+SEM_CRIT	86.0	56.8	85.2	61.6	62.4	**54.2**	**61.7**	59.3
EP+SEM_PRIN	84.0	62.0	83.2	61.2	67.6	52.1	56.9	**61.6**
Llama 4 Scout								
Zero Shot	82.0	66.4	57.2	66.8	49.9	51.9	47.9	49.2
Few Shot	90.7	74.4	69.6	66.4	61.3	**54.5**	58.1	**58.4**
EP_CRIT	**92.0**	77.6	**82.8**	**70.0**	**61.5**	51.2	**59.1**	57.2
SEM_CRIT	84.7	62.8	66.8	63.2	48.9	47.8	48.8	51.3
EP+SEM_CRIT	91.3	**78.4**	**82.8**	68.8	57.8	51.9	55.9	57.6
o4-mini								
Zero Shot	**92.7**	87.6	89.2	62.0	50.4	49.3	51.1	51.6
Few Shot	91.3	90.0	**91.6**	**66.8**	60.0	49.7	63.9	59.3
EP_CRIT	**92.7**	80.8	89.2	64.8	60.6	50.5	68.1	**60.7**
SEM_CRIT	92.0	69.6	88.8	60.4	48.0	48.2	48.9	50.9
EP+SEM_CRIT	91.3	**90.4**	90.8	61.2	**61.5**	**52.4**	**68.3**	57.6

Table 1: Agent accuracy across datasets and models. We use EP, SEM, and EP+SEM to denote episodic, semantic, and combined memory. Suffixes _CRIT and _PRIN indicate critique- or principle-based entries. Results on preference datasets are averaged across all users. For each model and dataset, the highest score is **bolded** and the second-highest is underlined.

of memory-augmented learning. The first four datasets are more fact-oriented and show minimal to no improvement from our learning pipeline when using gpt-4o-mini. Although we significantly outperform the zero-shot baseline on the Multi-Condition Ranking task, failing to exceed the few-shot baseline suggests that improvements stem from few-shot prompting rather than from the insights themselves.

In contrast, three out of four preference-based datasets exhibit clear gains. Incorporating insights yields over 10% improvement on all three preference datasets, except for *Book Pref*. For interpretability, results are averaged over the three users per dataset, though notable variation exists across users (see Figure 2 for per-user results).

Episodic memory generally outperforms semantic memory, indicating that the model benefits more from a few specific examples than from a summary of the entire training set—suggesting limitations in the quality of the semantic summaries. Still, combining the specificity of episodic memory with the generalizations of semantic memory sometimes yields the strongest performance. Principle-based methods tend to perform well in domains where knowledge is more clustered and are often more effective with episodic memory. This may be because summarizing over principles can produce overly generic representations that lack the actionable specificity needed for effective decision-making.

5.4 Results with Different Models

We observe substantial variation in how the three backbone models respond to memory-augmented learning. In principle, the critic and performance agents may use different models; however, to constrain experimental complexity, we use the same model for both. Additionally, we select one representative strategy per memory type for each model.

LLaMA 4 Scout (Meta, 2025) exhibits lower baselines on fact-oriented datasets but stronger few-shot performance, particularly on preference data—an opposite trend compared to gpt-4o-mini. We attribute these differences to model size and variations in pretraining emphasis. Llama 4 Scout is a mixture-of-experts model with 17B active parameters and 109B total parameters. This opens different opportunities for memory-augmentated learning: we observe strong gains on NFCorpus and modest improvements on Multi-Condition Ranking and PubMed. On preference datasets, while seeing improvements over zero-shot, it shows little to no gain. Semantic mem-

ory alone is consistently uncompetitive, whereas episodic memory often yields the best performance. Adding semantic memory to episodic memory occasionally helps, notably in Multi-Condition Ranking.

O4-mini (OpenAI, 2025), a reasoning-focused model, achieves the highest baselines on fact-oriented tasks—even in the zero-shot setting, leaving minimal room for improvement via utilizing memory. Only `EP_SEM_CRIT` provides marginal gains over this strong baseline. However, on preference datasets, combining episodic and semantic memory consistently outperforms both zero- and few-shot baselines, suggesting this model is particularly capable of integrating complementary signals from both memory types.

Experiment (Training Percentage)	Steam Pref	Book Pref	Anime Pref	Movie Pref
Baselines				
Zero-shot	52.8	52.0	47.9	49.9
Few-shot	55.8	50.9	51.1	53.2
`EP_CRIT`				
25%	61.6	49.5	55.7	56.1
50%	62.9	51.3	54.9	57.3
75%	**64.1**	53.4	**57.6**	57.1
100%	62.7	**53.8**	54.4	**57.7**
`EP+SEM_CRIT`				
25%	57.8	48.5	55.1	57.1
50%	60.1	51.2	55.6	56.8
75%	59.7	50.6	58.9	**60.5**
100%	**62.4**	**54.2**	**61.7**	59.3

Table 2: Accuracy with varying size of training dataset on preference datasets using `gpt-4o-mini`. For each strategy, the highest score is **bolded**.

5.5 Dataset Size Scaling

One of the strengths of this learning approach is how it is able to significantly improve performance over baseline with a very small amount of labeled data, especially using episodic memory. To test how much data is required, we ran `gpt-4o-mini` through the pipeline again on each of the preference datasets, using only 25%, 50%, and 75% of the original training data. Both the `EP` and `EP+SEM` methods begin to show improvement at 25%, with accuracy continuing to increase as more data is incorporated into memory. Methods utilizing semantic memory are more affected by the lack of training data, as this can lead to lower-quality summary-level insights. Performance frequently begins leveling off between 75% and 100%, implying that we are reaching saturation with these datasets.

5.6 Suggestibility

In memory-augmented agentic learning, it is crucial not only to generate the best possible insight (critique or principle) for inclusion in memory, but also to ensure that the model is actually receptive to it, i.e., that it can be "persuaded" by the insight. This receptivity, which we term *suggestibility*, is influenced by a compound of factors: the model architecture, the nature of the task, and the format in which the memory is represented.

To better quantify this phenomenon, we define a *suggestibility* metirc S, which captures the difference in an agent's performance when given a best-effort insight versus when given an intentionally misleading one (generated by flipping the ground-truth label). Formally,

$$S = \frac{1}{|D|} \sum_{x_i \in D} \mathbb{1}\left[\text{PA}(x_i \mid \text{Ins}(x_i, y_i)) = y_i\right] - \frac{1}{|D|} \sum_{x_i \in D} \mathbb{1}\left[\text{PA}(x_i \mid \text{Ins}(x_i, \neg y_i)) = y_i\right]$$

where PA denotes the performance agent, Ins refers to the insight generation agent (which may produce a critique or principle), and D is the evaluation dataset. Note that in real-world settings, the true label y_i is not available to either PA or Ins; thus, this metric represents an idealized or "cheating" scenario, using artificially constructed best and adversarial insights for controlled experimentation.

To explore how different components affect a model's suggestibility, we report S across five experimental conditions, varying the context provided to the performance agent. As shown in Table 3, `X` indicates the presence of the question, `Y` denotes inclusion of the ground-truth label, and `Crit/Prin` represent whether a critique or principle is included.

Model suggestibility exhibits strong dependence on task characteristics. Fact-based datasets tend to produce low suggestibility, consistent with expectations. These models are fine-tuned to resist misinformation and are generally reluctant to give confidently incorrect answers. A notable exception is the PubMed dataset, where the technical complexity of medical queries appears to introduce enough ambiguity for insights to meaningfully influence the model's output. In contrast, preference-based datasets reveal high levels of suggestibility in nearly all conditions, except when only a principle is presented without an accompanying answer.

	MMLU	Multi-Condition Ranking	NFCorpus	PubMed	Steam Preference	Book Preference	Anime Preference	Movie Preference
`XY`	16.7	45.6	6.4	98.4	100.0	100.0	100.0	100.0
`XY+Crit`	48.7	98.4	40.0	99.6	100.0	100.0	100.0	100.0
`XY+Prin`	16.0	40.4	4.8	98.4	99.4	99.7	98.2	99.5
`X+Crit`	58.0	100.0	70.8	93.2	100.0	99.8	100.0	100.0
`X+Prin`	0.7	-2.0	0.0	8.0	10.6	17.6	18.0	13.8

Table 3: Suggestibility scores across datasets averaged across users. `X` represents the question, `Y` denotes the presence of an answer, and `Crit/Prin` represent whether additional insights are present.

This, too, is expected: the model lacks specific knowledge about the user's preferences and will often adopt whichever answer it is guided toward.

Models demonstrate a marked increase in suggestibility when explanations are provided. This is evidenced by higher S in the `XY+Crit` condition relative to both `XY` alone and `XY+Prin`. Furthermore, providing the true label substantially improves suggestibility in principle-based settings (compare `XY+Prin` with `X+Prin`), but yields smaller improvements—or even slight declines—in critique-based settings (compare `XY+Crit` with `X+Crit`). This may be because critiques already contain pointwise assertions, whereas principles tend to express more general, future-oriented guidance. Detailed accuracy values for the performance agent under both true and false insight conditions are included in the appendix (Table 4).

6 Related Work

Agentic Memory Recent LLM-based agent research has focused on memory management challenges due to context length limitations. The predominant approach is retrieval-based augmentation (RAG), using embedding similarity for memory retrieval. Memories range from simple input/output copies to complex structures: Reflexion (Shinn et al., 2023) stores agent self-reflections, Voyager (Wang et al., 2024) maintains reusable agent-created tools, and Generative Agents (Park et al., 2023) employs a two-tier system of event streams and higher-level reflections.

Fine-tuning While fine-tuning is a well-established way of improving a model's performance in a specific area (Dodge et al., 2020), it presents challenges such as: extensive labeled data requirements (Vieira et al., 2024), catastrophic forgetting (Luo et al., 2025), computational expense (Hu et al., 2022), and inapplicability to closed-source models.

In-Context Learning In-context learning treats models as black boxes, adjusting inputs to influence outputs (Dong et al., 2024). Simple prompt modifications like appending "Let's think step by step" can significantly improve performance (Kojima et al., 2022). Few-shot learning enhances results by providing question-answer examples (Brown et al., 2020). Reflection-based approaches, where models reason over feedback about their decisions, enable autonomous improvement (Shinn et al., 2023; Yao et al., 2023). However, most research focuses on feedback from simulated environments (Wang et al., 2024), with limited exploration of other feedback mechanisms.

7 Conclusion

In summary, we present a memory-augmented framework for enabling LLM agents to continuously learn from supervised signals and structured critiques. By modeling both semantic and episodic memory and introducing reflective insights in the form of critiques and principles, our approach enhances generalization without modifying model parameters. Empirical results demonstrate substantial performance gains and reveal the importance of memory structure and critique quality. Our proposed metric, suggestibility, offers a new lens for understanding how models internalize feedback. This work highlights the potential of reflective memory as a lightweight, interpretable, and extensible mechanism for continual adaptation in LLMs.

Limitations

Our analysis has centered on the accuracy of different agentic learning strategies, but design choices also impact computational cost and the ability to incorporate ongoing supervision. Semantic memory typically requires greater training-time computation due to summarization or distillation, whereas episodic memory simply stores past experiences

with minimal processing. At inference time, however, semantic memory offers more readily applicable knowledge, while episodic memory relies on retrieval quality. This trade-off suggests that the optimal strategy may depend on the size of the supervised dataset and the frequency of inference—semantic memory may be better suited for frequent inference under sparse supervision, while episodic memory may be preferable when supervision is abundant and retrieval is reliable.

An additional interesting direction for exploring model suggestibility is to disentangle how much a model's behavior changes due to genuinely incorporating supervised signals into its internal beliefs versus merely adapting its responses to please the user. In our empirical study, we observed that models exhibited higher suggestibility scores when critiques were attributed to the user, compared to when the same critiques were believed to originate from the model itself or another model. This suggests that the perceived source of feedback plays a significant role in how seriously the model treats the signal, opening up opportunities to better understand and guide belief formation in interactive learning systems.

The agentic learning tasks we explore primarily involve constraints on classification tasks, particularly in question answering and preference prediction. We observed significant performance variations across different domains, and it remains unclear whether these findings will generalize beyond the tasks and domains studied. Additionally, we found that performance generally improved with larger training sizes. Although this improvement appeared to plateau, larger datasets are needed to better understand how these techniques scale beyond the sampled training sets used in this study.

To reduce computational costs, we used the same model for both the performance agent and the critic agent. The impact of using different models within the memory-augmented learning setup remains unexplored. It is possible that combining models of varying capacities could yield many of the benefits of a more powerful model, particularly in terms of insight generation.

Ethics Statement

This research on memory-augmented learning for large language model agents raises several important ethical considerations that we wish to acknowledge.

Though our suggestibility work was focused on how the model's instruction-following ability varied with dataset, this kind of approach could also be used to more efficiently jailbreak models to spread misinformation. Future work should be careful to avoid developing tools to improve the suggestibility of models to the point that they spread harmful misinformation.

We also recognize that improved adaptation capabilities may exacerbate existing biases in these agents. Because the insights are generated by the agent itself, even with feedback from the labeled data, it could cause the agent to reinforce its preconceptions about the world, which may perpetuate harmful stereotypes. Future work should explore safeguards to identify and mitigate such bias amplification.

References

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2024. Teaching large language models to self-debug. In *The Twelfth International Conference on Learning Representations*.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024.

CRITIC: Large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2025. An empirical study of catastrophic forgetting in large language models during continual fine-tuning.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

Meta. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

OpenAI. 2025. Introducing openai o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/.

Manas Parashar. Movie recommendation system. https://www.kaggle.com/datasets/parasharmanas/movie-recommendation-system.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Pouya Pezeshkpour and Estevam Hruschka. 2025. Multi-conditional ranking with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2863–2883, Albuquerque, New Mexico. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 8634–8652. Curran Associates, Inc.

Theodore Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas Griffiths. 2024. Cognitive architectures for language agents. *Transactions on Machine Learning Research*. Survey Certification.

Tamber. Steam video games. https://www.kaggle.com/datasets/tamber/steam-video-games/data.

Cooper Union. Anime recommendations database. https://www.kaggle.com/datasets/CooperUnion/anime-recommendations-database.

Inacio Vieira, Will Allred, Séamus Lankford, Sheila Castilho, and Andy Way. 2024. How much data is enough data? fine-tuning large language models for in-house translation: Performance evaluation across multiple dataset sizes.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2024. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2015. Memory networks. Publisher Copyright: © 2015 International Conference on Learning Representations, ICLR. All rights reserved.; 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19724–19731.

Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification.

A Implementation Details

A.1 RAG Implementation

We used blevlabs/stella_en_v5 as our encoder model and FAISS as our vector database. Similarity was based purely on the encodings of the questions in each dataset.

Though we did experiment with fine-tuning an encoder model to increase the separation of the classes in each dataset (for example, bought vs not-bought games for each Steam user) in embedding space, we did not see significant improvements in performance.

A.2 Dataset: Additional Details

NFCorpus: The original NFCorpus data source associated each article with many papers with varying degrees of separation, which we transformed into this pairwise setup by choosing one paper at the closest and furthest level of separation possible for each paper. Sampled 500 shortest combinations of articles and papers to avoid context-length issues.

Steam and Book Preference: Due to limitations in the number of games/books per user, the training and test sizes are smaller for these datasets than others. The train-test split percentage was maintained at 50%.

Steam User 1679: 104 samples in train set
Steam User 3188: 129 samples in train set
Steam User 6839: 116 samples in train set
Book User 63: 218 samples in train set
Book User 123: 183 samples in train set
Book User 2642: 206 samples in train set

A.3 Models

Default hyperparameters were used for all models. OpenAI models were queried through the OpenAI API, Llama 4 Scout was queried through the fireworks.ai API.

B Prompts

Critique Generation

```
User: {Question}
Agent: {PA Initial Prediction}
User: The correct answer is {Ground Truth Answer}. Explain why this is the correct answer, following the following JSON format
{
  correct_answer: correct_answer,
  local_reason: Specific reasons why this answer is correct in this particular case.,
  global_reason: General reasons why this answer is correct that can be applied to other questions.
}.
Respond only with JSON.
```

Principle Generation

```
Your task is to identify trends in data to improve your ability to make correct predictions in the future.

For example in examples:
  {Question} {Answer}

Identify one and only one guiding principle explaining why the given answers are correct. If there are no obvious connections between the questions, give a principle for the first example only.
A guiding principle may be a fact, a pattern, a preference, or anything else that will help you answer questions like these in the future. With that in mind, principles should be focused on information that you do not already know. It should be very specific and not generic advice.
Respond only with the principle, nothing else.
```

Semantic Memory Generation

```
Your job is to summarize a set of self-critiques made by some agent as they perform different instances of their task. For each instance you will be shown the output of the agent, followed by the critiques made
```

by the agent after they were told the correct answer. Distill those critiques into a helpful summary of advice to the agent, paying particular attention to instances where the agent outputs an incorrect answer. Produce your output in a form that can be used directly as instructions to the agent. You should summarize the key points in these critiques. Be precise and concise. Do not repeat yourself.
For example in train_set:
{Question} {Answer} {Critique}

Performance Agent with Semantic Memory

{Question}
Here is some helpful advice that will help you make your decision: {Summary}

Performance Agent with Episodic Memory

For example in examples:
User: {Example Question}
Agent: {PA Initial Prediction}
User: {Critique Generation Prompt}
Agent: {Critique}
User: Here is your final question, make sure to learn from your past mistakes! {Question}

Performance Agent with Episodic and Semantic Memory

For example in examples:
User: {Example Question}
Agent: {PA Initial Prediction}
User: {Critique Generation Prompt}
Agent: {Critique}
User: Here is your final question, make sure to learn from your past mistakes! {Question}
Also, here is some additional advice to guide your response: {Summary}

C Examples

Example Critique Summary (NFCorpus)

1. **Focus on Relevance**: Always choose the reference that directly relates to the subject matter of the article. Look for references that support the main claims made in the article.
2. **Identify Key Themes**: Ensure that the reference paper closely aligns with the key themes discussed in the article, such as specific health effects, mechanisms of action, or relevant population studies.
3. **Avoid General Topics**: Select references that do not deviate into unrelated topics. If one reference discusses foundational knowledge or statistics that do not support the article's claims, it's likely not the correct choice.
4. **Highlight Specific Effects**: When discussing studies, emphasize specific effects or outcomes that are directly addressed in the article. Look for quantitative data or direct correlations that would affirm the article's claims.
5. **Example Comparison**: When there are multiple choices, conduct a clear comparison between them. If one reference explicitly discusses the same variables outlined in the article, that should be favored.
6. **Review Findings**: When evaluating findings from referenced studies, ensure they corroborate the arguments or recommendations presented in the article. This can include discussing potential risks, benefits, or mechanisms.
7. **Address Opinions and Recommendations**: When the article discusses guidelines or opinions (such as on health recommendations), favor references that critique or analyze these points directly.
8. **Check for Clinical Relevance**: In clinical or scientific discussions, emphasize studies that provide empirical evidence that can be tied back to practical outcomes related to the topic of the article.
9. **Nutritional Context**: In discussions around diet, ensure the references speak to the nutritional context being examined, such as the impact of specific foods on health, rather than unrelated dietary patterns.
10. **Summarizing Connections**: When concluding which reference is correct, clearly summarize why the chosen reference aligns best with the article's content. Discuss how it supports or expands upon the article's key points.
By following these instructions, you will ensure that your references are relevant and provide strong support for the claims made in the articles you analyze.

Example Principle Summary (Anime Preference)

User 1635 exhibits a strong preference for anime that features strong character devel-

opment, emotional depth, and complex narratives. Their ratings indicate enjoyment of series such as "Clannad: After Story," "Gintama," "Higurashi no Naku Koro ni," and "Kimi ni Todoke," while they tend to rate simpler or less character-driven titles lower, such as "Skip Beat!" and "Nyan Koi!".
They are particularly drawn to genres that blend action, adventure, and psychological themes, often rating these series highly (8/10 or above). Titles like "Naruto," "Psycho-Pass," and "Fairy Tail" resonate well with them, while they show less enthusiasm for slice-of-life or lighter narratives.
User 1635 also tends to favor critically acclaimed or popular series, indicating a preference for well-regarded storytelling and character arcs.
Overall, their anime preferences reflect a consistent inclination towards emotionally engaging and character-driven narratives, while they are less likely to enjoy works that lack depth or complexity.

D Additional Results

On the fact-based datasets in Table 4, the accuracy on the False trials is much higher than on the preference datasets. This supports the theory that the low suggestibility on these datasets is primarily due to the models being fine-tuned specifically to avoid giving disinformation.

In Figure 2 there is significant variation between users on the same dataset. It is expected that some users will have more eclectic interests and be more difficult to predict, but it is notable how much the best strategy differs between each user. On the Steam Preference dataset, despite the overall improvements being similar, the best approach varies from `EP+SEM_CRIT`to `EP_PRIN`to `EP+SEM_PRIN`.

Experiment	MMLU	Multi-Condition Ranking	NFCorpus	PubMed	Steam Preference	Book Preference	Anime Preference	Movie Preference
`XY True`	96.0	92.0	89.6	99.2	100.0	100.0	100.0	100.0
`XY False`	79.3	46.4	83.2	0.8	0.0	0.0	0.0	0.0
`XY+Crit True`	98.7	100.0	98.0	100.0	100.0	100.0	100.0	100.0
`XY+Crit False`	50.0	1.6	58.0	0.4	0.0	0.0	0.0	0.0
`XY+Prin True`	94.7	93.6	86.8	99.2	99.7	99.7	99.9	99.9
`XY+Prin False`	78.7	53.2	82.0	0.8	0.3	0.0	1.7	0.4
`X+Crit True`	99.3	100.0	99.6	96.8	100.0	100.0	100.0	100.0
`X+Crit False`	41.3	0.0	28.8	3.6	0.0	0.2	0.0	0.0
`X+Prin True`	90.0	60.0	83.6	58.8	57.2	61.4	58.4	56.9
`X+Prin False`	89.3	62.0	83.6	50.8	46.6	43.8	40.4	43.1

Table 4: Raw accuracy in suggestibility analysis, showing specific accuracy scores when given the True or False answer. X represents the question to answer, Y represents whether the answer to the question is given or not, and Crit/Prin represent whether additional insights are present. Results on preference datasets are averaged across all users.

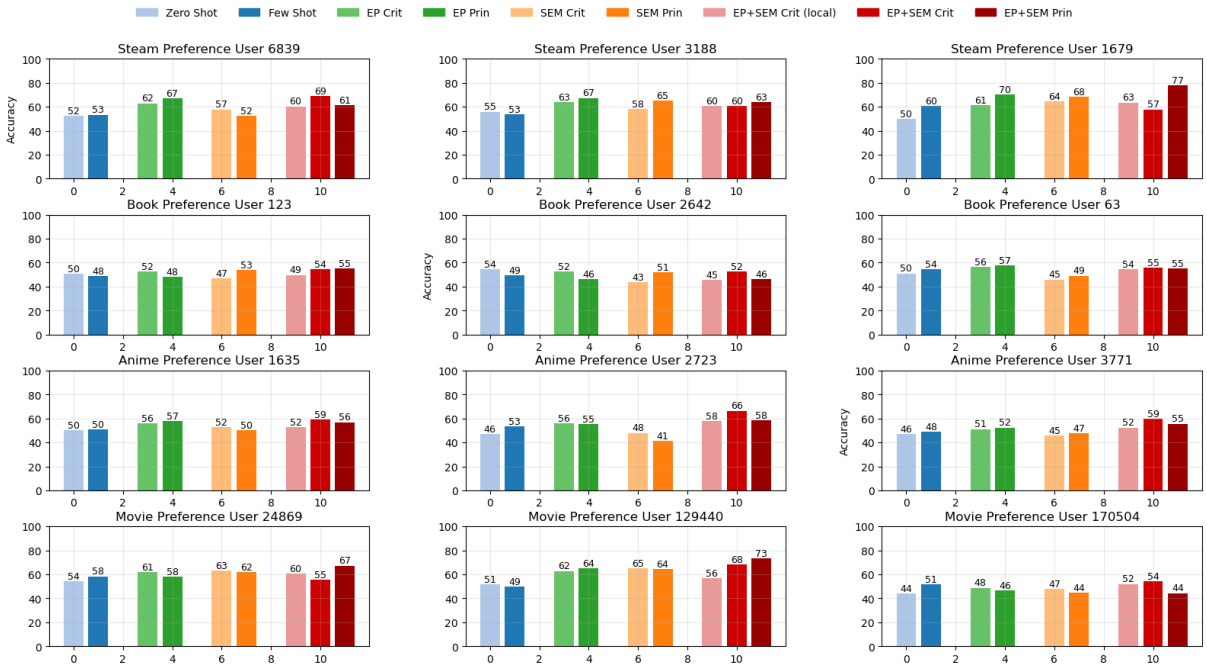

Figure 2: Preference data accuracy results by user. We use `EP`, `SEM`, and `EP+SEM` to denote episodic, semantic, and combined memory. Suffixes `_CRIT` and `_PRIN` indicate critique- or principle-based entries.