Hypothesis-Driven Reasoning for Large Language Models

Aakash Kumar Agarwal[1*†], **Moyuru Yamada**[2†]

[1]Indian Institute of Technology, Bombay, India,
[2]Fujitsu Research of India, Bangalore, India
yamada.moyuru@fujitsu.com

Abstract

This paper tackles the fundamental failure of Large Language Models (LLMs) to solve new tasks when prompted with a sufficient, yet overly complex, set of multi-modal episodes. This failure stems from the model's inability to distill underlying patterns from the noisy experiences. We propose Hypothesis-Driven Reasoning (HDR), a framework that enhances LLM reasoning by building an explicit semantic memory—a set of hypotheses induced from the multi-modal episodes. HDR employs a two-stage pipeline. It first extracts potential factors from the episodes and then iteratively refines hypotheses by generate-verify loop with the factors. We first empirically demonstrates this failure and the potential of sematic memory, showing that oracle hypotheses can boost accuracy from 35.3% to 92.0% on a novel task we designed. We then evaluate our HDR, achieving near-oracle performance and significantly outperforming baselines, especially on smaller models. This paper validates a shift from unstructured in-context recall to explicit knowledge abstraction for robust reasoning.

Introduction

The grand quest of artificial intelligence is to create agents that, like humans, learn from a stream of experiences and generalize to novel situations and tasks. Central to this endeavor is the concept of memory (Wu et al. 2025; Zhang et al. 2024). For Large Language Models (LLMs), the prompt context serves as a form of short-term working memory and the context window is regarded as its memory size. LLMs have a large working memory (An et al. 2024), and it is widely assumed that enriching this memory with more examples (i.e., past episodes) should enhance their reasoning capabilities (Brown et al. 2020). In this paper, however, we report a fundamental limitation of this notion. Using a diagnostic task we designed to probe rule-inference, we found that simply filling a LLM's short-term memory with raw episodic data leads to a large degradation in performance as the episodes become more complex. In contrast, providing a few pieces of semantic memory—abstract hypotheses distilled from the past episodes—results in a dramatic accuracy improvement. This findings highlight the significance of inducing semantic knowledge from discrete episodes.

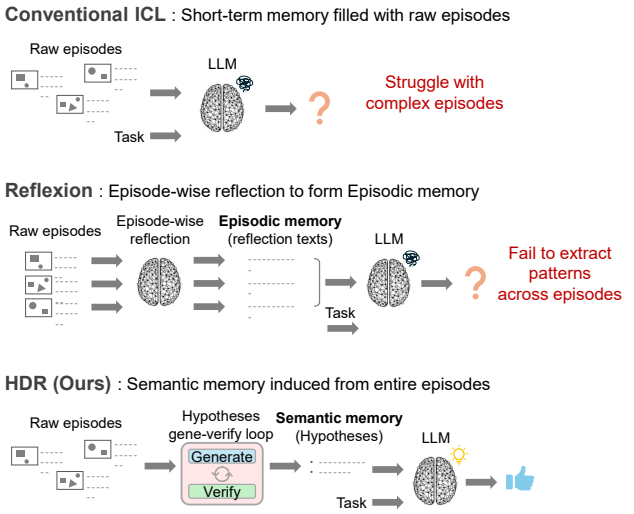

Figure 1: Standard In-Context Learning (ICL) performance degrades as raw episodes become more complex, whereas our Hypothesis-Driven Reasoning (HDR) induces a semantic memory from the episodes, enhancing reasoning capability. Reflexion creates a episodic memory on a sample-by-sample basis, failing to capture the patterns across episodes.

This challenge has spurred a surge of research into equipping LLMs with long-term memory. Pioneering efforts in this domain primarily focused on overcoming context window limitations, particularly for long-running tasks like extended dialogues (Maharana et al. 2024). These systems typically treat memory as a passive datastore, storing a history of raw episodes and retrieving relevant chunks when needed. While effective at extending conversational history, these methods do not perform active abstraction or learning from the stored memories. More recently, a new wave of research has begun to explore more active forms of memory (Xu et al. 2025; Zhong et al. 2024; Shinn et al. 2023), processing and transforming individual experiences into more structured forms on a sample-by-sample basis. This can be seen as creating a type of episodic or procedural memory.

*Work performed during an internship at FRIPL.
†These authors contributed equally.

Some pioneers (Li and Li 2024; Kim et al. 2023; Wang and Chen 2025) have also highlighted the semantic memory as a critical component for real-world applications. However, first, the effect of semantic memory on complex reasoning has been insufficiently explored in the context of modern LLMs. Second, the crucial challenge of synthesizing a global semantic memory from the entire collection of disparate multi-modal experiences remains unaddressed.

To fill these critical gaps, we propose *Hypothesis-Driven Reasoning (HDR)*, a novel framework designed explicitly to build a semantic memory for LLM reasoning. HDR mimics the cognitive process of memory consolidation, whereby disconnected short-term experiences are transformed into long-term knowledge. We employ a multi-modal large language model (MLLM) as a hypothesis generator to form an iterative generate-and-verify loop. First, the MLLM extracts potential factors over the multi-modal episodes. Then, it iteratively generates hypotheses—an explicit patterns stated in natural language and verifies generated hypotheses against available evidence, pruning flawed claims. Through this iterative refinement, HDR distills a noisy sea of raw episodic data into a powerful semantic memory.

Our primary contributions are as follows:

- We demonstrate and quantify the impact of semantic memory on a simple yet challenging multi-modal experience-transfer task.
- We propose Hypothesis-Driven Reasoning (HDR), a novel framework designed to autonomously build an explicit semantic memory by synthesizing hypotheses across multi-modal experiences.
- We empirically validate our framework, showing that the semantic memory formed by HDR dramatically improves LLM capabilities.

Related Work

This paper primarily focuses on enhancing reasoning capabilities through the use of hypotheses. To situate this contribution, we first review the literature on complex reasoning with memory, highlighting the largely unexplored potential of semantic memory in the context of LLMs. We then survey works on hypothesis generation with LLMs, positioning our framework as the pioneer to tackle this challenge in the multi-modal domain.

Complex Reasoning with Memory

The ability of LLMs to perform complex reasoning is fundamentally tied to their capacity to leverage information provided as context. Initial research demonstrated that LLMs can perform in-context learning (ICL), where performance on a new task improves as more examples are provided in the prompt (Brown et al. 2020), which acts as a form of short-term memory (Wu et al. 2025). A significant body of work has since focused on optimizing this short-term memory, for instance, by carefully selecting the most informative examples (or episodes) to present (Zhang, Feng, and Tan 2022; Purohit et al. 2025). However, this paradigm is inherently limited by the finite context window and does not support learning across the episodes.

To overcome these limitations, recent studies have explored equipping agents with various forms of long-term memory. Retrieval-Augmented Generation (RAG) (Lewis et al. 2020) emerged as a pivotal development. RAG equips LLMs with a form of long-term memory by connecting them to an external, passive data store, typically a vector database. Following the great success of RAG, researchers have extended RAG mechanism to address the limitations on long-term dialogues (Maharana et al. 2024). They store the chat history in a memory bank and summarize a chat session (Zhong et al. 2024) or link different sessions (Xu et al. 2025). They mainly focus on retrieving relevant past interactions from the memory for generating a proper response. We focus on the failure of reasoning caused not by a lack of data, but by the complexity of the sufficient episodes provided. Reflexion (Shinn et al. 2023) and its successors (Wang et al. 2024; Kim et al. 2024) are a key step in this direction. It allows an agent to verbally reflect on a episode to identify mistakes and improve its performance in subsequent attempts. While this represents a significant advance towards agents that learn from experience, the learning remains fundamentally local, reflection is conducted separately for each individual episode to form episodic memory.

Some pioneers have highlighted an importance of semantic memory and integrated it into their system (He et al. 2024; Wang and Chen 2025). They extract abstract knowledge from the textual data and use it as semantic memory to improve the performance. However, the impact of semantic memory on challenging reasoning tasks, particularly those requiring multi-modal experience transfer, has not been evaluated enough.

Hypotheses Generation with LLMs

A promising pathway to building the semantic memory posited above is through automated hypothesis generation. Recently, with modern LLM capability, a new attempts in automatic inductive reasoning has emerged. They feed observations as prompts into the LLM, sampling multiple hypotheses that explains the observations. IDEA (He et al. 2024) uses a simple CoT prompt to generate hypotheses. HypoGeniC (Zhou et al. 2024) introduces a reward function to inform the exploitation-exploration tradeoff in the update process for generating high-quality hypotheses with LLM. Another attempt (il Lee et al. 2024) is first to generate distinct concepts and feed them into the LLM separately for hypothesis generation (w/o iterative refinement). One-stage iterative hypothesis generation is also recently introduced (Qiu et al. 2024).

However, the aforementioned approaches have been developed and validated exclusively within the textual domain. The challenge of abstracting principles from noisy, high-dimensional multi-modal signals—where text is grounded in vision—has remained an open frontier. HDR is a novel framework designed to bridge this divide with LLMs. Unlike the existing method, we introduce a novel two-stage iterative pipeline to build semantic memory from multi-modal episodes. Note that some papers visualize samples in MiniARC dataset (Kim et al. 2022) while they are originally textual data.

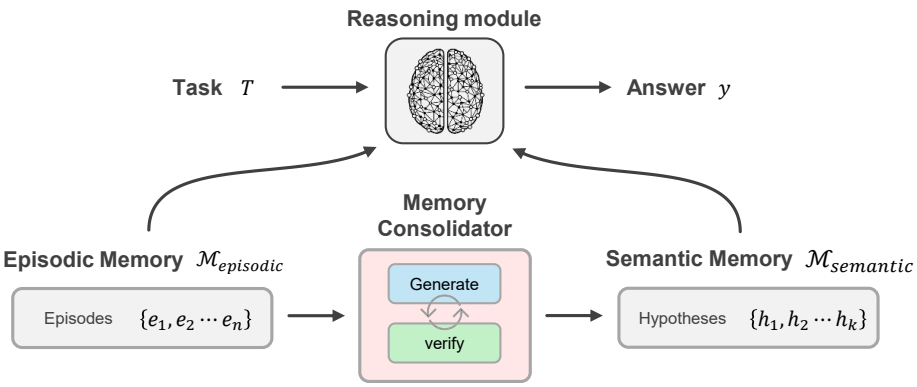

Figure 2: An overview of Hypothesis-Driven Reasoning (HDR) framework, which enhances reasoning capability with semantic memory induced from episodic memory. Memory Consolidator generates hypotheses from raw episodes and stores them in the semantic memory.

Hypothesis-Driven Reasoning

We introduce *HDR*, Hypothesis-Driven Reasoning that enhances the LLM reasoning capability with semantic memory induced from episodic memory. The overview is shown in Fig. 2. We also provide a simplified algorithm demonstrating how the HDR operates in Algorithm 1.

Overview

Let $\mathcal{M}_{\text{episodic}} = \{e_1, e_2, \ldots, e_n\}$ denotes the episodic memory, where each e is a multi-modal episode. The episode e_i contains an object o_i (e.g., an image of colored shape) and a associated label l_i (e.g, category, description, or reflection). HDR aims to derive an answer y for a task T with a semantic memory $\mathcal{M}_{\text{semantic}}$ induced from $\mathcal{M}_{\text{episodic}}$. The semantic memory $\mathcal{M}_{\text{semantic}}$ is composed of interpretable hypotheses $\mathcal{H} = \{h_1, h_2, \ldots, h_k\}$, such that these hypotheses capture latent rules underlying over the episodes.

HDR consists of the episodic memory $\mathcal{M}_{\text{episodic}}$, the semantic memory $\mathcal{M}_{\text{semantic}}$, a memory consolidator MC, and a reasoning module R.

$$\mathcal{M}_{\text{semantic}} = MC(\mathcal{M}_{\text{episodic}})$$

$$y = R(T, \mathcal{M}_{\text{episodic}}, \mathcal{M}_{\text{semantic}})$$

The memory consolidator employs two-stage pipeline: *factor extraction* and *iterative hypothesis refinement* for generating reliable hypotheses from the episodes.

Factor Extraction

Given episodic memory $\mathcal{M}_{\text{episodic}}$, we begin by extracting candidate *factors* (e.g., attributes such as color and position or textual keywords) that may explain the observed patterns over the episodes. This factor set $\mathcal{F} = \{f_1, \ldots, f_m\}$ forms the basis for hypothesis generation. We incorporate this crucial step for capturing patterns in visual data based on observations from our preliminary experiments. We initialize the $\mathcal{F}$ as $\mathcal{F}_0$ which is empty. We iteratively extract new factors and refine all the factor set $\mathcal{F}$ with a multi-modal LLM (MLLM) as a factor extractor FE until the factor set stops changing or a predetermined number of iteration K.

$$\mathcal{F}_i = FE(\mathcal{M}_{\text{episodic}}, \mathcal{F}_{i-1})$$

Iterative Hypothesis Refinement

This step generates the reliable hypotheses by iteratively and alternatively performing *hypothesis generation* and *hypothesis verification* until the designated number of times N. The pipeline is shown in Fig. 3.

Hypothesis Generation With the extracted factors $\mathcal{F}$, the MLLM is prompted as a hypothesis generator HG to produce candidate hypotheses $\mathcal{H} = \{h_1, h_2, \ldots\}$.

$$\mathcal{H}_i = HG(\mathcal{M}_{\text{episodic}}, \mathcal{F}, \mathcal{H}_{i-1})$$

Initially, the hypotheses is set to empty and previous hypotheses are not taken into account. After generating initial hypotheses, this step also refines previous hypotheses. The hypothesis generator attempts to generalize from observed patterns to candidate semantic rules. For example:

- *Yellow squares are anomalous.*
- *All red shapes are anomalies.*

Hypothesis Verification Once a set of candidate hypotheses $\mathcal{H}$ has been generated, we evaluate each hypothesis for consistency with the episodic memory $\mathcal{M}_{\text{episodic}}$. We employ MLLM as a hypothesis verifier HV, which verifies whether the hypothesis is valid or not:

$$\mathcal{M}_{\text{semantic}} = \{h_i \in \mathcal{H} \mid HV(\mathcal{M}_{\text{episodic}}, h_i) = \texttt{valid}\}$$

The full set of hypotheses is presented together in a single prompt alongside the episodic memory. The model returns a *discrete validity judgment*—`valid` or `invalid`—for each hypothesis, indicating whether it is consistent with the episodes. We store only hypotheses explicitly marked as `valid` by the verifier and discard the other invalid hypotheses. This strict filtering mechanism improves the reliability of the semantic memory and reduces error propagation in downstream inference.

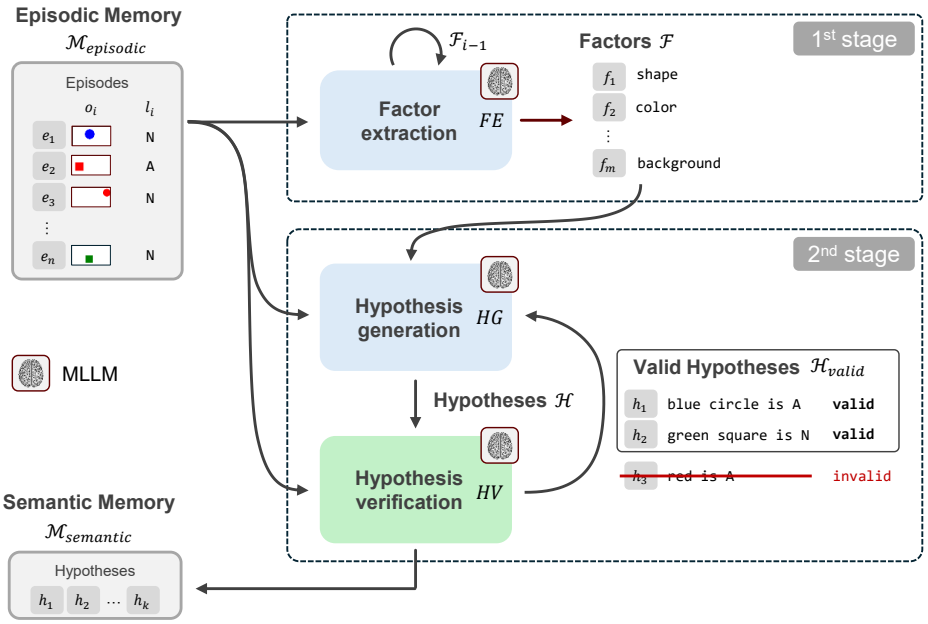

Figure 3: Memory consolidator pipeline, which employs two-stage pipeline: *factor extraction* and *iterative hypothesis refinement* for generating reliable hypotheses from the multi-modal episodes.

Algorithm 1: Hypothesis-Driven Reasoning (HDR)

Input: Task T, MLLM R, $\mathcal{M}_{\text{episodic}} = \{e_1, \ldots, e_n\}$
Output: Answer y

Two-stage Memory Consolidator
Stage 1: Factor Extraction
Initialize $\mathcal{F}_0 \leftarrow \emptyset$
for $i = 1$ **to** K **do**
  $\mathcal{F}_i \leftarrow FE(\mathcal{M}_{\text{episodic}}, \mathcal{F}_{i-1})$
  if $\mathcal{F}_i = \mathcal{F}_{i-1}$ **then**
    break
  end if
end for
$\mathcal{F} \leftarrow \mathcal{F}_K$
Stage 2: Iterative Hypothesis Refinement
Initialize $\mathcal{H}_0 \leftarrow \emptyset$
for $j = 1$ **to** N **do**
  $\mathcal{H} \leftarrow HG(\mathcal{M}_{\text{episodic}}, \mathcal{F}, \mathcal{H}_{j-1})$
  $\mathcal{H}_{\text{valid}} \leftarrow \{h \in \mathcal{H} \mid HV(h, \mathcal{M}_{\text{episodic}}, \mathcal{F}) = \texttt{valid}\}$
  $\mathcal{H}_i \leftarrow \mathcal{H}_{\text{valid}}$
end for
$\mathcal{M}_{\text{semantic}} \leftarrow \mathcal{H}_N$
Reasoning with Memory
$y = R(T, \mathcal{M}_{\text{episodic}}, \mathcal{M}_{\text{semantic}})$
return y

$\mathcal{M}_{\text{episodic}}$: Episodic memory
$\mathcal{M}_{\text{semantic}}$: Semantic memory
FE: Factor extractor, HG: Hypothesis generator,
HV: Hypothesis verifier

Inference

At test time, the inference model is presented with a novel task T composed of a query q and a test object o_t. HDR conditions the MLLM not only on the episodic memory but also on the distilled semantic memory, i.e., hypotheses:

$$\text{Input}_t = \mathcal{M}_{\text{episodic}} \cup \mathcal{M}_{\text{semantic}} \cup T$$

The semantic memory enhances the model's reasoning capability by providing foundation for multi-step reasoning on noisy episodes.

Experiments

This section first introduces a novel task called *Experience-Transfer task*. We then describes baseline methods and implementation details.

Experience-Transfer Task

The Experience Transfer task is designed to evaluate a model's ability to extract semantic knowledge from episodic experiences and apply it compositionally to solve more complex problems (Fig. 4). This extends beyond conventional in-context learning (ICL) by requiring the model to construct a generalizable semantic memory that supports zero-shot reasoning in unfamiliar contexts.

Formally, the inference model is first provided with an episodic memory $\mathcal{M}_{\text{episodic}} = \{e_1, e_2, \ldots, e_n\}$ where each e is a multi-modal observation. The observation e_i contains an object o_i (e.g., a colored shape within an image), and $l_i \in \{\texttt{normal}, \texttt{anomaly}\}$ is its associated anomaly label. Each episode defines a local regularity (e.g., color-shape rules) that governs anomaly assignment within that episode.

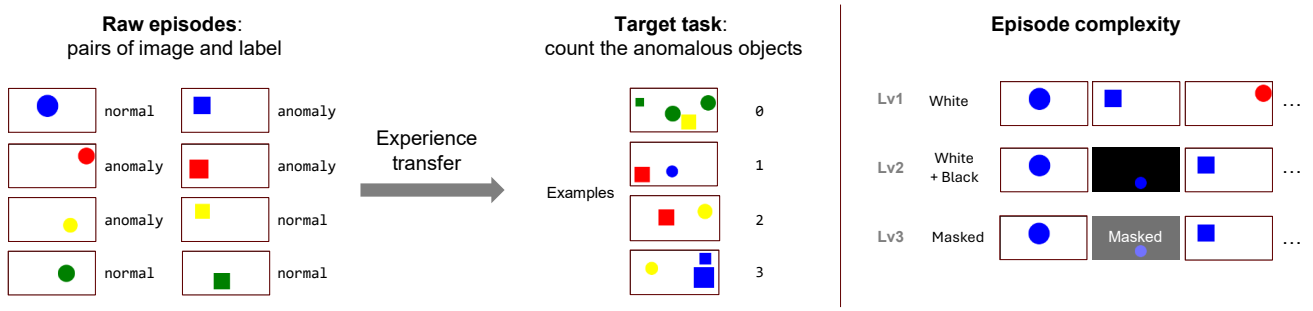

Figure 4: Experience-Transfer task, which required to understand underlying patterns over episodes for solving a target task. We design three levels of complexity regarding to the episodes.

At test time, the model is given a target task T composed of a query q and test scene o_t. The scene consists of $m_t \in \{2, 3, 4\}$ objects, sampled from the object in the episodic memory. The scenes introduce combinations of previously observed objects not seen during episodic training.

The task is to predict the number of anomalous objects in the test scene. The query q describe the task, i.e., "*Count total number of anomalous objects in the image*".

$$\hat{y}_t = \sum_{j=1}^{m_t} \hat{l}_{t,j}, \quad \hat{l}_{t,j} \in \{0, 1\}$$

where the number of anomalous objects is 0 - 4.

This task is simple yet challenging because it requires LLMs to perform multi-step reasoning over textual and visual modalities.

We designed three levels of complexity:

- **Level 1:** All images have a white background. This results in 8 unique combinations (2 shapes $\times$ 4 colors).
- **Level 2:** Images features both white and black backgrounds. This results in 16 unique combinations (2 shapes $\times$ 4 colors $\times$ 2 backgrounds).
- **Level 3:** We carefully masked 4 objects out of 16 in Level 2. Masking increases the complexity of the reasoning task while still retaining just enough examples to make rule induction possible.

The label of each object (normal or anomaly) were determined based on predefined rules given in Table 1. These rules are not explicitly provided to the reasoning model, requiring it to infer them from episodes.

Data Synthesis

To ensure our task evaluates in-context reasoning rather than recalled knowledge, we generate a synthetic dataset. This approach allows us to embed unique patterns that are deliberately absent from the LLM's pre-training corpus. We also intentionally avoid using complex objects such as real-word entities (e.g., 'blue sedan car') to prevent our evaluation from being confounded by an unrelated failure mode (e.g., visual grounding). Our objective is not to test LLM's object recognition capabilities, but to isolate the more foundational skill of reasoning over latent rules.

Shape	Color	Background	Label
Square	Red	White	Anomaly
Circle	Red	White	Anomaly
Square	Green	White	Normal
Circle	Green	White	Normal
Square	Blue	White	Anomaly
Circle	Blue	White	Normal
Square	Yellow	White	Normal
Circle	Yellow	White	Anomaly
Square	Red	Black	Normal
Circle	Red	Black	Normal
Square	Green	Black	Anomaly
Circle	Green	Black	Anomaly
Square	Blue	Black	Normal
Circle	Blue	Black	Anomaly
Square	Yellow	Black	Anomaly
Circle	Yellow	Black	Normal

Table 1: Anomaly labeling rules. Rules flip when the background is black (e.g., red becomes normal, green becomes anomaly, etc.). Four objects are masked in Lv. 3 (gray).

Episode data We created synthetic raw episode images by placing a single colored geometric shape—either a circle or a square—on a rectangular canvas of size 200×100 pixels. The color of the shape (red, blue, yellow, or green) and the background (white or black) were systematically varied to ensure full coverage of all combinations. Objects were placed randomly within the canvas, and shape sizes were sampled uniformly from a fixed range (25–50 pixels).

Test Data The test set comprised 150 images, evenly distributed across object counts: 50 images each with 2, 3, or 4 objects. Each image included randomly placed non-overlapping objects, sampled from the same set of shapes, colors, and backgrounds used in raw episodes. The ground-truth anomaly count was determined using the same rules applied as raw episodes. In the Level 3, some test images also included object types absent from raw episodes, requiring indirect inference via learned semantic patterns.

Methods	CoT	Lv. 1			Lv. 2			Lv. 3		
		GPT	Qwen	Gemma	GPT	Qwen	Gemma	GPT	Qwen	Gemma
Episodic memory		66.0	75.3	18.0	35.3	42.7	20.0	40.7	20.7	22.0
Semantic memory		**91.3**	**90.7**	**56.7**	**92.0**	**88.3**	31.3	**76.0**	**66.0**	37.3
Episodic + Semantic		88.7	87.3	46.0	90.7	76.7	**34.7**	68.0	62.0	**42.0**
Episodic memory	✓	86.7	75.3	52.0	77.3	45.3	34.0	55.3	19.3	40.7
Semantic memory	✓	**91.3**	**92.7**	76.0	90.7	79.3	60.0	**68.7**	**68.7**	56.0
Episodic + Semantic	✓	86.7	91.3	**86.3**	**94.7**	**80.7**	**76.0**	66.0	59.3	**61.3**

Table 2: Results on Experience-Transfer task with different memories. Mean accuracy (%) across three repetitions. We test three models of different sizes, GPT-4o, Qwen2.5-32B, and Gemma-12B. We use all the raw episodes as the episodic memory and oracle hypotheses as the semantic memory. CoT denotes Chain-of-Thought prompting.

Note that our dataset is simple yet scalable, allowing us to assess the LLM on more complex scenarios.

Metric

We report QA accuracy on our Experience-Transfer task.

Baselines

We test following baselines to evaluate the effect of semantic memory and our HDR framework.

Episodic Memory. The model receives only raw episodes without any form of abstraction. It must reason over these unstructured samples to solve novel test queries.

$$\mathcal{M}_{\text{semantic}} = \emptyset, \quad \text{Input}_t = \mathcal{M}_{\text{episodic}} \cup T$$

Semantic Memory. The model relies only on oracle hypotheses $\mathcal{H}_{gt}$ (no episodic memory).

$$\mathcal{M}_{\text{semantic}} = \mathcal{H}_{gt}, \quad \text{Input}_t = \mathcal{M}_{\text{semantic}} \cup T$$

Episodic Memory + Reflexion. Reflection on each individual episode is generated as proposed in (Shinn et al. 2023) and serves as the semantic memory.

$$\mathcal{M}_{\text{semantic}} = \{r_{\text{ind},0}, r_{\text{ind},1}, \ldots\}, \quad r_{\text{ind},i} = \text{Reflect}(e_i)$$
$$\text{Input}_t = \mathcal{M}_{\text{episodic}} \cup \mathcal{M}_{\text{semantic}} \cup T$$

Episodic Memory + Reflexion on all episodes. A single global reflection is generated from the full set of episodes instead of individuals. This reflection serves as a natural language summary of observed patterns, aiding the model with a compressed but informal abstraction.

$$\mathcal{M}_{\text{semantic}} = \{r_{\text{all}}\}, \quad r_{\text{all}} = \text{Reflect}(\mathcal{M}_{\text{episodic}})$$
$$\text{Input}_t = \mathcal{M}_{\text{episodic}} \cup \mathcal{M}_{\text{semantic}} \cup T$$

Episodic Memory + Direct Hypothesis. Hypotheses are directly generated by MLLM and fed into the model along with all the raw episodes.

Episodic Memory + Simple two-stage. Factors are first extracted from the episodes and then hypotheses are generated only once based on the extracted factors.

These baselines allow us to isolate the contribution of structured abstraction. While raw episodic memory is fragile in the face of complexity, even unverified hypotheses provide a scaffold for reasoning. Note that most existing methods assume the task stored in the episodic memory and the target task are the same, which means they cannot be directly applied to our Experience-Transfer task.

Implementation Details

Raw episodes were fed in random order using a fixed seed. We employed Gemini 2.5 Flash as our factor extractor, hypothesis generator, and hypothesis verifier. Inference was performed using GPT-4o (gpt-4o-2024-08-06), Qwen2.5-32B, and Gemma3-12B. All models were run with temperature = 0 and a fixed seed for deterministic output. However, GPT-4o still exhibited nondeterministic behavior, so all experiments were repeated 3 times and mean QA accuracy is reported. We used LangChain and LangGraph for orchestration. Open-source models (Qwen2.5-32B, Gemma3-12B) were run on a single NVIDIA A100 GPU; proprietary models were accessed via API. The number of iterations were determined via a few preliminary experiments (K=2, N=3).

Results and Analysis

This section first presents potential of semantic memory on Experience-Transfer task with oracle hypotheses. We then compare the performance of our Hypotheses-Driven Reasoning (HDR) with other baselines.

Potential of Semantic Memory

We evaluate accuracy on our Experience-Transfer task with different memories, i.e., episodic memory (feeding all the episodes), semantic memory (feeding all oracle hypotheses), and both episodic memory and semantic memory. We test three models of different sizes (∼200B, 32B, and 12B). Table 2 clearly shows that semantic memory significantly improves accuracy across all settings (all the models, with CoT and without CoT). Notably, for GPT-4o on Lv. 2, performance increased from 35.3% to 92.0%. The intuition behind this result is that providing common knowledge for the tasks to reach the answer can avoid failures stemming from brittle, ad-hoc multi-step reasoning on noisy episodes (i.e., non-necessary episodes for the specific test sample can be the noise and degrades the performance). Semantic memory and episodic memory with CoT prompting achieves the highest accuracy throughout the experiment, while the model is struggling in Level 3 where some episodes are masked. Inspired by this result, we propose Hypothesis-Driven Reasoning (HDR).

Methods	Lv. 1			Lv. 2			Lv. 3		
	GPT	Qwen	Gemma	GPT	Qwen	Gemma	GPT	Qwen	Gemma
Episodic + Semantic	86.7	91.3	86.3	94.7	80.7	76.0	66.0	59.3	61.3
Reflextion (Individual)	76.7	86.7	34.0	59.3	30.0	26.0	52.7	19.3	38.0
Reflextion (All)	**84.0**	79.3	32.7	71.3	23.3	26.7	53.3	19.3	34.0
HDR (Ours)	**84.0**	**94.0**	**82.7**	**84.7**	**78.0**	**80.0**	**60.0**	**60.7**	**59.3**

Table 3: Performance comparison on Experience-Transfer task. Mean accuracy (%) across three repetitions for GPT. Our Hypothesis-Driven Reasoning (HDR) archives significantly better performance on most settings.

Models	Lv. 1	Lv. 2	Lv. 3	Ave.
Gemini-2.5 Flash	84.0	**84.7**	**60.0**	**76.2**
GPT-4o	87.3	84.0	45.3	72.2
Qwen-2.5-32B	89.3	78.7	39.3	69.1
Gemmma3-12B	**90.7**	43.3	53.3	62.4

Table 4: Effect of the memory consolidation models. We test different models for memory consolidation, while keeping GPT-4o as the inference model on test samples.

Performance of HDR

We compare our HDR with other baseline methods as shown in Table 3. Our method archives significantly better performance on 8 out of 9 settings. This improvement is observed across all models and particularly significant when the model size is small, e.g., 26.0% vs 80.0% in Level 2 with Gemma. Performance of our HDR nearly reaches the results achieved using the oracle hypotheses. In some cases, our method even performs better due to randomness.

Ablation Studies

We also conduct two ablation studies as shown in Table 4 and Table 5.

Effect of Memory Consolidation Models We employ different models for memory consolidation and evaluate accuracy on our task, while keeping GPT-4o as the inference model. Table 4 shows that Gemini achieves the best overall performance in hypothesis generation. Interestingly, other models outperform Gemini in Level 1.

Effect of Factor Extraction and Iterative Refinement We perform an ablation study to evaluate the effectiveness of the key features of our HDR. We manually assess the generated hypotheses and compute the F1 score for each method. Table 5 demonstrates superior performance of our HDR, generating all the true hypotheses. Compare to that Direct hypothesis shows miserable results, failing to identify the key factors and generating the wrong hypotheses. Simple two-stage shows excellent performance in Level 1 and 2, while it only induces a part of hypotheses in Level 3. These results demonstrate the importance of both factor extraction and iterative refinement incorporated in our HDR framework.

Methods	FE	IR	Lv. 1	Lv. 2	Lv. 3
Direct hypothesis			0.67	0.0	0.0
Simple two-stage	✓		1.0	1.0	0.36
HDR (Ours)	✓	✓	**1.0**	**1.0**	**1.0**

Table 5: F1 score of hypothesis generation across different methods. FE and IF denotes the factor extraction and the iterative hypothesis refinement, respectively.

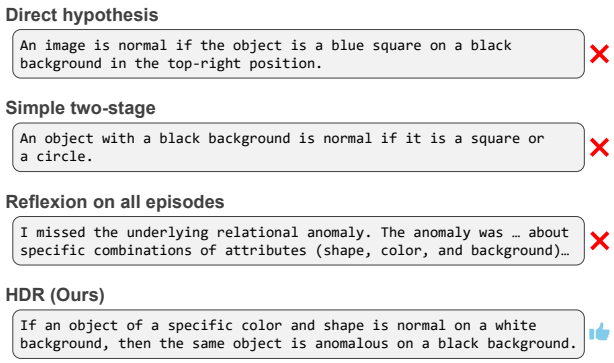

Figure 5: Examples of generated hypotheses in Level 3.

Qualitative Analysis

We show some examples in Fig. 5. Our HDR captures the underlying complex patterns, while the others fail.

Conclusion

We first demonstrated that while LLMs struggle to reason with raw episodic memory, their capabilities are dramatically unlocked by explicit semantic memory. Inspired by this key insight, we introduced Hypothesis-Driven Reasoning (HDR), a novel framework that automatically distills semantic hypotheses from multi-modal experiences via a generate-and-verify loop. Experiments compellingly show that our HDR framework significantly outperforms baselines and achieves near-oracle performance. This work validates a crucial shift from unstructured in-context learning to explicit, knowledge-driven reasoning, encouraging new avenues for building AI agents that learn by actively forming and reasoning with semantic memory.

References

An, C.; Zhang, J.; Zhong, M.; Li, L.; Gong, S.; Luo, Y.; Xu, J.; and Kong, L. 2024. Why Does the Effective Context Length of LLMs Fall Short? *ArXiv*, abs/2410.18745.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T. J.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; teusz Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. *ArXiv*, abs/2005.14165.

He, K.; Zhang, M.; Yan, S.; Wu, P.; and Chen, Z. 2024. IDEA: Enhancing the Rule Learning Ability of Large Language Model Agent through Induction, Deduction, and Abduction. In *Annual Meeting of the Association for Computational Linguistics*.

il Lee, K.; Koh, H.; Lee, D.; Yoon, S.; Kim, M.; and Jung, K. 2024. Generating Diverse Hypotheses for Inductive Reasoning. *ArXiv*, abs/2412.13422.

Kim, S.; Phunyaphibarn, P.; Ahn, D.; and Kim, S. 2022. Playgrounds for Abstraction and Reasoning. In *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)*.

Kim, S. H.; iunn Ong, K. T.; Kwon, T.; Kim, N.; Ka, K.; Bae, S.; Jo, Y.; won Hwang, S.; Lee, D.; and Yeo, J. 2024. Towards Lifelong Dialogue Agents via Timeline-based Memory Management. In *North American Chapter of the Association for Computational Linguistics*.

Kim, T.; Cochez, M.; François-Lavet, V.; Neerincx, M.; and Vossen, P. 2023. A machine with short-term, episodic, and semantic memory systems. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 9459–9474. Curran Associates, Inc.

Li, J.; and Li, J. 2024. Memory, Consciousness and Large Language Model. *ArXiv*, abs/2401.02509.

Maharana, A.; Lee, D.-H.; Tulyakov, S.; Bansal, M.; Barbieri, F.; and Fang, Y. 2024. Evaluating Very Long-Term Conversational Memory of LLM Agents. *ArXiv*, abs/2402.17753.

Purohit, K.; V, V.; Bhattacharya, S.; and Anand, A. 2025. Sample Efficient Demonstration Selection for In-Context Learning. In *Forty-second International Conference on Machine Learning*.

Qiu, L.; Jiang, L.; Lu, X.; Sclar, M.; Pyatkin, V.; Bhagavatula, C.; Wang, B.; Kim, Y.; Choi, Y.; Dziri, N.; and Ren, X. 2024. Phenomenal Yet Puzzling: Testing Inductive Reasoning Capabilities of Language Models with Hypothesis Refinement. In Kim, B.; Yue, Y.; Chaudhuri, S.; Fragkiadaki, K.; Khan, M.; and Sun, Y., eds., *International Conference on Representation Learning*, volume 2024, 1452–1481.

Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K. R.; and Yao, S. 2023. Reflexion: language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Wang, G.; Xie, Y.; Jiang, Y.; Mandlekar, A.; Xiao, C.; Zhu, Y.; Fan, L.; and Anandkumar, A. 2024. Voyager: An Open-Ended Embodied Agent with Large Language Models. *Transactions on Machine Learning Research*.

Wang, Y.; and Chen, X. 2025. MIRIX: Multi-Agent Memory System for LLM-Based Agents. *ArXiv*, abs/2507.07957.

Wu, Y.; Liang, S.; Zhang, C.; Wang, Y.; Zhang, Y.; Guo, H.; Tang, R.; and Liu, Y. 2025. From Human Memory to AI Memory: A Survey on Memory Mechanisms in the Era of LLMs. *ArXiv*, abs/2504.15965.

Xu, W.; Liang, Z.; Mei, K.; Gao, H.; Tan, J.; and Zhang, Y. 2025. A-MEM: Agentic Memory for LLM Agents. *ArXiv*, abs/2502.12110.

Zhang, Y.; Feng, S.; and Tan, C. 2022. Active Example Selection for In-Context Learning. *ArXiv*, abs/2211.04486.

Zhang, Z.; Bo, X.; Ma, C.; Li, R.; Chen, X.; Dai, Q.; Zhu, J.; Dong, Z.; and Wen, J.-R. 2024. A Survey on the Memory Mechanism of Large Language Model based Agents. *ArXiv*, abs/2404.13501.

Zhong, W.; Guo, L.; Gao, Q.; Ye, H.; and Wang, Y. 2024. MemoryBank: Enhancing Large Language Models with Long-Term Memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17): 19724–19731.

Zhou, Y.; Liu, H.; Srivastava, T.; Mei, H.; and Tan, C. 2024. Hypothesis Generation with Large Language Models. In *Proceedings of EMNLP Workshop of NLP for Science*.