StateTree: Enhancing Long-Term Dialogue Reasoning via Reinforcement Learning

Naen Xu[1,2*] **Wanqing Cui**[2*] **Yibo Hu**[2†] **Shixin Hong**[2]
Hengyu An[1] **Meiguang Jin**[2] **Junfeng Ma**[2] **Tianyu Du**[1‡]
[1]Zhejiang University [2]Taobao & Tmall Group of Alibaba
{xunaen, anhengyu, zjradty}@zju.edu.cn
{cuiwanqing.cwq, boxuan.hyb, hongshixin.hsx}@taobao.com
{meiguang.jmg, jack.majf}@taobao.com

Abstract

Large language models deployed as personalized assistants must reason over long, evolving interaction histories. However, in long-term dialogue reasoning, relevant evidence is scattered across sessions, preferences may be revised over time, and standard long-context training fails to address these challenges under data scarcity and prohibitive computational costs. We propose StateTree, a data-driven RL method that constructs a challenging auxiliary task from scarce dialogues with verifiable ground truth. StateTree augments multi-session dialogues with a tree-structured path-tracing task: key-value records are embedded across sessions to form a binary tree. Solving the task requires the model to traverse from root to leaf by retrieving records across sessions and comparing timestamps to resolve branches, then recover the hidden target question among distractor leaves. We apply curriculum RL training progressively increasing tree depth and introduce a compositional variant whose edges carry step-level reasoning fragments, training the model to compose partial cues into coherent queries. Trained on 10K-token contexts, StateTree generalizes to 128K tokens without full-length RL costs and exhibits capabilities including cross-session retrieval, temporal reasoning, knowledge update, and compositional multi-hop reasoning. StateTree outperforms both SFT and RL-based baselines while preserving short-context general reasoning. StateTree-7B achieves gains up to +23.60% on LongMemEval (128k), and StateTree-14B reaches 59.00% accuracy on LongMemEval, surpassing QwenLong-L1-32B (51.00%)[1].

1 Introduction

Large Language Models (LLMs) are increasingly deployed as personalized assistants for writing [1, 2], recommendation [3, 4, 5, 6], and consultation [1, 7], while supporting long-term, multi-session dialogues [8]. However, LLMs struggle to leverage personal knowledge accumulated over extended interaction histories [9, 10], leading to degraded accuracy and reduced user satisfaction [11]. While memory-augmented systems [9, 12] improve personalization via compression, indexing, and retrieval over chat histories, they still rely on the backbone LLM's reasoning ability. Meanwhile, most long-context research targets static documents [13, 14], treating dialogue history as flat text and failing to adapt to evolving user personas [15, 16]. Long-term dialogue reasoning presents three challenges:

- **Data Scarcity ($\mathbb{C}1$):** Existing multi-session datasets such as LoCoMo [15] contain only 10 dialogues (1540 QA pairs), and LLM-synthesized dialogues cannot guarantee answer correctness.

[*] Equal contribution.
[†] Project leader.
[‡] Corresponding author.
[1] Code and dataset are available at `https://github.com/bluedream02/StateTree`.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

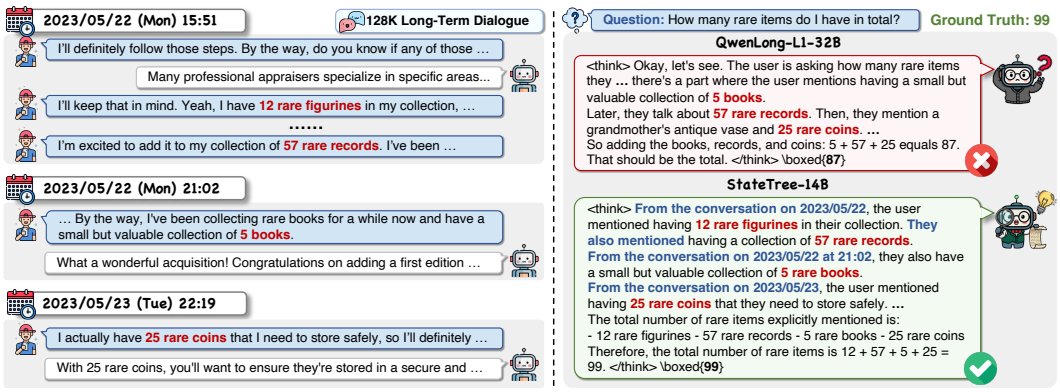


Figure 1: Model trajectories in multi-session dialogue reasoning. (i) QwenLong-L1-32B gets lost across sessions and misses information. (ii) The StateTree-trained model shows cross-session retrieval, temporal reasoning, knowledge update, and multi-hop reasoning, improving reasoning reliability.

- **Complex Dialogue Reasoning** ($\mathbb{C}2$)**:** Unlike static documents, long-term dialogues are non-stationary: evidence is scattered across sessions with timestamps, and preferences may be revised later. Answering question requires chaining multiple pieces of evidence, yet LLMs frequently fail at cross-session retrieval, temporal reasoning, knowledge update, and multi-hop composition.
- **Prohibitive Computational Cost** ($\mathbb{C}3$)**:** Training at near-target context lengths [17, 18] incurs prohibitive compute costs and risks degrading short-context and general reasoning abilities [19, 20].

While synthetic trajectories for SFT [21, 22, 23] mitigate data scarcity ($\mathbb{C}1$), models trained on static traces show limited improvement and often fail to generalize to out-of-distribution datasets. Reinforcement Learning (RL) [24, 25] offers an alternative, yet its effectiveness in long-term dialogue is constrained by the lack of challenging training data with ground truth ($\mathbb{C}1$) and prohibitive costs of long interaction chains ($\mathbb{C}3$). This leads to our central question: *How can we elicit robust* ***long-term dialogue reasoning*** *($\mathbb{C}2$) from extremely* ***scarce data*** *($\mathbb{C}1$) without* ***prohibitive computational overhead*** *of full-context RL training ($\mathbb{C}3$)?*

In response, we introduce **StateTree**, a data-driven RL pseudo-task that constructs challenging synthetic auxiliary tasks embedded within authentic dialogues. Our key insight is that multi-session dialogue reasoning can be reframed as a path-search problem: relevant evidence forms a navigable structure with temporal constraints and hierarchical dependencies. These patterns can be taught through structured synthetic tasks that provide verifiable ground truth. Concretely, StateTree embeds key-value records into long-term dialogues, transforming limited data into challenging training instances ($\mathbb{C}1$). The same key is placed across multiple sessions with different timestamps, and the linked records form a binary tree. The model must trace a path from root to leaf by comparing timestamps and following the most recent record at each branch point, until the leaf reveals the target question ($\mathbb{C}2$). We apply a curriculum RL strategy that progressively increases tree depth and reasoning complexity; its final stage introduces a *Compositional StateTree* whose edges carry step-level reasoning fragments that the model must aggregate into a coherent query. By training at contexts of 10K tokens, StateTree avoids the prohibitive cost of full-length RL ($\mathbb{C}3$).

Our experiments show that StateTree substantially improves Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct and Qwen3-8B on long-term dialogue reasoning. StateTree-14B achieves an average accuracy of 60.91% on LoCoMo, surpassing much larger baselines including QwenLong-L1-32B. Trained on contexts of 10K tokens, StateTree models generalize to long-term conversation reasoning on much longer 128K benchmarks (LongMemEval, PersonaMem-128k), and StateTree preserves short-context reasoning abilities. Notably, a qualitative comparison reveals that StateTree-trained models exhibit four emergent reasoning behaviors (cross-session retrieval, temporal reasoning, knowledge update, and multi-hop composition) that directly align with the four challenges encoded in the StateTree design, confirming that the synthetic task successfully imparts targeted reasoning patterns that generalize beyond the training distribution. Our main contributions are:

- We propose StateTree, a data-driven RL pseudo-task that augments scarce multi-session dialogues with tree-structured key-value records, transforming limited data into challenging training instances with verifiable ground truth to enhance long-term dialogue reasoning.

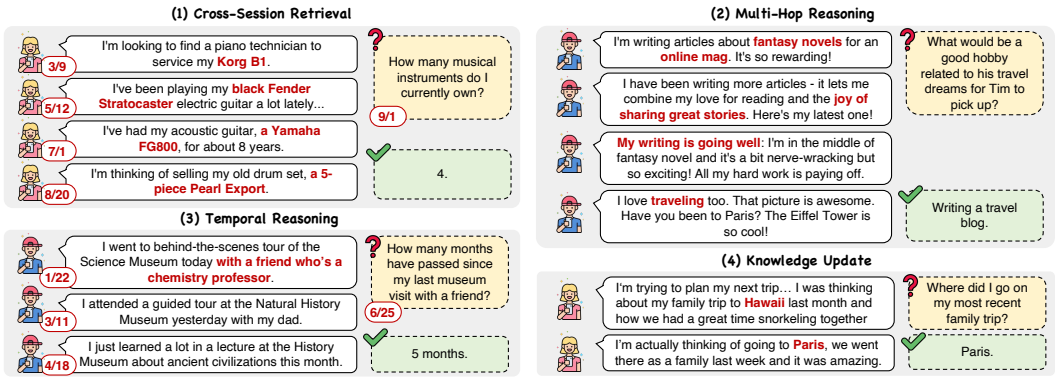


Figure 2: Examples of the challenges in multi-turn dialogue. For each example, we show the associated evidence statements on the left and the question with the answer on the right.

- We develop a curriculum RL strategy that progressively increases tree depth and introduces a compositional variant with step-level semantic fragments, enabling stable training and eliciting targeted reasoning behaviors.
- StateTree elicits emergent reasoning behaviors correspond to its task design. Despite training on only 616 examples at 10K tokens, it generalizes to 128K contexts without full-length RL costs, preserves short-context abilities, and matches or exceeds the performance of much larger models.

2 Related Work

Evaluating Long-term Multi-Session Dialogues. As LLMs are increasingly deployed as personalized assistants, understanding their ability to reason over extended interaction histories has become critical. A key finding is that models exhibit a "lost-in-the-middle" effect, with greater difficulty recalling information in the middle of the context [26]. [10] further show that when LLMs take a wrong turn in multi-turn dialogue, they fail to recover. Recently, evaluation has shifted toward more realistic dialogue interactions with sessions ranging from 10K to over 100K tokens. Representative benchmarks include LoCoMo [15], LongMemEval [16], and PersonaMem [27]. These evaluations reveal that even strong models struggle with long-term dialogue reasoning, motivating two lines of work: memory-augmented systems [28] and long-context reasoning methods.

Long-Context Reasoning. Memory-augmented systems [9, 12] externalize memory through compression, indexing, and retrieval [29, 30], but ultimately depend on the backbone LLM's ability to reason over retrieved context. Advanced long-context reasoning primarily utilizes synthetic-data SFT [31, 32] and RL [24], both of which are constrained by the biases of static synthetic traces or the base model's intrinsic reasoning capacity. For example, [31] proposes SEALONG, a self-improvement approach that samples multiple reasoning outputs and selects high-quality traces for SFT or preference alignment. QwenLong-L1 [33] applies RL to extend long reasoning trajectories up to 60K tokens. Inspired by needle-in-a-haystack [34], which measures models' retrieval ability in extremely long-text settings, LoongRL [35] inserts chains that hide the true question among distracting documents to support advanced long-context reasoning. While these approaches enhance long-context reasoning, they are designed for static documents rather than the non-stationary, multi-session dialogue setting.

3 StateTree

As shown in Figure 2, long-term interaction dialogues pose four challenges: (*i*) ***Cross-Session Retrieval***, which requires precisely locating sparse, relevant evidence scattered across numerous sessions [15]; (*ii*) ***Multi-Hop Reasoning***, necessitating bridging multiple retrieval steps across sessions and synthesizing the aggregated fragments into a final answer [15, 16]; (*iii*) ***Temporal Reasoning***, which involves comparing timestamps across sessions to ground events in chronology [16]; and (*iv*) ***Knowledge Update***, which requires recognizing that the same topic may be discussed at different times and that the most recent information should take precedence [36].

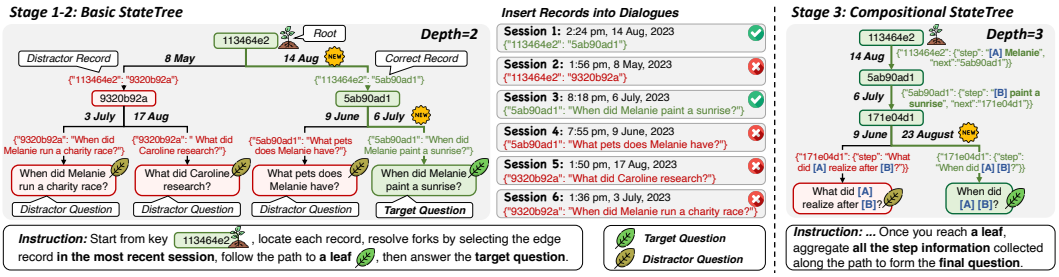


Figure 3: The StateTree data construction pipeline. **Left:** Basic StateTree embeds key-value records across sessions into a tree; the model traverses root-to-leaf to recover the target question. **Right:** Compositional StateTree augments edges with step-level fragments aggregates into the final question.

Table 1: Mapping between multi-turn dialogue challenges and StateTree design choices.

Challenge	**Design in StateTree**
Cross-Session Retrieval	Edge records distributed evenly across S sessions, preventing positional shortcuts
Multi-Hop Reasoning	D-level tree traversal; compositional StateTree requires aggregating step fragments into the question
Temporal Reasoning	At each fork, compare session timestamps to determine chronological ordering
Knowledge Update	Correct edge is placed in a newer session than the distractor, mirroring real recency preference

To overcome these challenges, we propose StateTree, a data-driven RL pseudo-task that constructs challenging synthetic auxiliary tasks embedded within authentic dialogues, targeting the capabilities required for long-term dialogue understanding and providing verifiable ground truth (see Figure 3). It combines (i) a data construction pipeline (Section 3.1) for challenging task generation; and (ii) GRPO with structured curriculum RL training for progressive skill acquisition through increasing tree depth and reasoning complexity, together with a combined reward function using exact match reward to provide clear optimization signals that directly target answer correctness (Section 3.2). These pillars collectively provide a manageable learning path for stable, data-efficient LLM training. Table 1 summarizes how each challenge is addressed by a specific design choice in StateTree.

3.1 Data Construction Pipeline

The core challenge in training LLMs for long-term dialogue reasoning is the absence of tasks that demand multi-step reasoning over extended dialogue histories. To address this, we design a data-driven RL pseudo-task with a data construction pipeline that embeds enhanced reasoning challenges directly within authentic multi-session dialogue data.

Base Data. We build upon multi-session dialogue datasets (e.g., LoCoMo [15]) where each sample consists of multiple dialogue sessions between two speakers, each annotated with a timestamp. Each sample is paired with factual QA pairs whose answers can be derived from the dialogue content.

Task Overview. The core of our data construction is *StateTree*, a complete binary tree of depth D built around a target question-answer pair (q_i, a_i). The target question q_i is hidden in one leaf node, while all edges are encoded as JSON key-value records scattered across dialogue sessions. Solving the task requires traversing from root to leaf to recover the hidden question and produce a_i.

Consider a multi-session dialogue $\mathcal{S} = \{s_1, \ldots, s_S\}$, where each session is annotated with a timestamp, paired with a set of questions $\mathcal{Q}$ and their ground-truth answers. For a target question $q_i \in \mathcal{Q}$ with answer a_i, we construct a StateTree $\mathcal{T}$ and embed its edges as JSON records throughout the sessions, yielding an augmented dialogue $\mathcal{S}'$. Each training instance is then a tuple $(\mathcal{S}', \mathcal{P}, a_i)$: the model receives $\mathcal{S}'$ together with an instruction prompt $\mathcal{P}$ specifying the root key k_{root} and traversal rules, and must find q_i and produce a_i. The design follows three principles, each targeting a specific challenge: **(i)** the original question q_i is *hidden* and edge records are distributed across all S sessions, so the model must perform *cross-session retrieval* to locate them; **(ii)** at each fork, the correct edge record resides in a *more recent session* than the distractor, forcing *temporal reasoning* and *knowledge update*; **(iii)** the model must navigate D levels of forks to reach the leaf, then answer the recovered question, exercising *multi-hop reasoning*. We employ two tree formats: a *Basic StateTree* whose leaves directly contain questions, and a *Compositional StateTree* whose edges carry step-level reasoning fragments that must be aggregated to recover the question.

Algorithm 1 Basic StateTree Construction.

Require: $\mathcal{S} = \{s_1, \ldots, s_S\}$, QA pairs $\mathcal{Q}$, depth D
Ensure: Augmented dialogue $\mathcal{S}'$, prompt $\mathcal{P}$, answer a_i

$(q_i, a_i) \sim \mathcal{Q}$
Build $\mathcal{T}$ (depth D, internal nodes V_{int}, leaves L); assign UUID key k_v to each node
Sample $\ell^* \in L$; set $q_{\ell^*} \leftarrow q_i$ and $q_\ell \sim \mathcal{Q} \setminus \{q_i\}$ for $\ell \neq \ell^*$
for each $v \in V_{\text{int}}$ with children u_1, u_2 **do**
  for $j \in \{1, 2\}$ **do**
    if $u_j \in V_{\text{int}}$ **then**
      $r_j \leftarrow \{k_v : k_{u_j}\}$
    else
      $r_j \leftarrow \{k_v : q_{u_j}\}$
    end if
  end for
  if v on root-to-ℓ^* path **then**
    $u_{\text{corr}} \leftarrow$ on-path child, $u_{\text{dist}} \leftarrow$ sibling
    Pick $s_a, s_b \in \mathcal{S}$ with timestamp$(s_b) >$ timestamp(s_a)
    Assign r_{corr} to s_b and r_{dist} to s_a
  else
    Assign r_1, r_2 to distinct sessions evenly
  end if
end for
Insert all records at sentence boundaries $\rightarrow \mathcal{S}'$
$\mathcal{P} \leftarrow$ PROMPT$(k_{\text{root}}$, traversal rules$)$;
return $(\mathcal{S}', \mathcal{P}, a_i)$

Algorithm 2 Compositional StateTree Construction.

Require: $\mathcal{S} = \{s_1, \ldots, s_S\}$, target (q_i, a_i), depth D, axes $\{A_1, \ldots, A_D\}$
Ensure: Augmented dialogue $\mathcal{S}'$, prompt $\mathcal{P}$, answer a_i

Build $\mathcal{T}$ (depth D, keys k_v); sample target leaf ℓ^*
Decompose q_i via LLM along axes, yielding target step $z_{(v,u)}$ for each edge on the root-to-ℓ^* path
Generate distractor steps for all other root-to-leaf paths via LLM, differing in ≥ 1 axis
for each internal node v with children u_1, u_2 **do**
  for $j \in \{1, 2\}$ **do**
    if $u_j \notin L$ **then**
      $r_j \leftarrow \{k_v : \{\texttt{step} : z_{(v,u_j)}, \texttt{next} : k_{u_j}\}\}$
    else
      $r_j \leftarrow \{k_v : \{\texttt{step} : z_{(v,u_j)}\}\}$
    end if
  end for
  if v on root-to-ℓ^* path **then**
    $u_{\text{corr}} \leftarrow$ on-path child, $u_{\text{dist}} \leftarrow$ sibling
    Pick $s_a, s_b \in \mathcal{S}$ with timestamp$(s_b) >$ timestamp(s_a)
    Assign r_{corr} to s_b and r_{dist} to s_a
  else
    Assign r_1, r_2 to distinct sessions evenly
  end if
end for
Insert all records at sentence boundaries $\rightarrow \mathcal{S}'$
$\mathcal{P} \leftarrow$ PROMPT$(k_{\text{root}}$, traversal and composition rules$)$;
return $(\mathcal{S}', \mathcal{P}, a_i)$

Basic StateTree. The Basic StateTree $\mathcal{T}$ is a complete binary tree of depth D. Each node is identified by a unique *key*—a randomly generated UUID string (e.g., `f391e945-...-9bbcecd4b90d`) carrying no semantic meaning, forcing the model to perform explicit lookup-and-follow operations rather than content-based shortcuts. Each of the $2^D - 1$ internal nodes has exactly two outgoing edges, encoded as JSON key-value records (e.g., {"k_i": "k_j"}). The 2^D leaf nodes each contain a natural-language question: exactly one holds the target question q_i; the rest hold distractor questions sampled from $\mathcal{Q} \setminus \{q_i\}$, ensuring all leaf questions are plausible and answerable from $\mathcal{S}'$, so the model cannot bypass tree navigation by simply selecting the most relevant-sounding question. There exists a unique correct path $k_{\text{root}} \rightarrow k_2 \rightarrow \cdots \rightarrow k_D \rightarrow q_i$ from the root to the target leaf. The tree contains $2(2^D - 1)$ edge records in total, all scattered across the dialogue sessions.

For each internal node on the correct path, its correct and distractor edge records are placed in *different* sessions, with the correct record always in the *more recent* one to force the model to perform *temporal discrimination* by comparing session timestamps to identify the most recent record. All edge records are inserted at sentence boundaries and distributed evenly across sessions to blend naturally into the dialogue flow. Algorithm 1 formalizes the construction of a Basic StateTree. The pipeline first builds a complete binary tree and scatters its edges across the dialogue sessions, then generates an instruction prompt that asks the model to traverse the tree and answer the recovered question. An example instance is provided in Appendix E.

Compositional StateTree. While the Basic StateTree trains cross-session retrieval, temporal reasoning, and knowledge update, the model never needs to *compose* information gathered along the path, which is a critical aspect of multi-hop reasoning. Thus, we introduce the *Compositional StateTree*, a D-level binary tree whose edges carry structured *step* information instead of bare key pointers. Concretely, given the target question q_i for a dialogue between two speakers, we use an LLM (prompt in Appendix C.4.1) to decompose q_i into a $2 \times 2 \times 2$ hierarchy along three semantic axes, which define the levels of a depth-3 binary tree with 8 leaves. Each edge carries a `step` field that provides a reasoning fragment at the corresponding level:

- **Person selection (Level 1):** which speaker is relevant to the dialogue, e.g., "`[A] Jon`"

- **Event category (Level 2):** which broad type of activity is involved, e.g., "`[B] expanding his studio's social media presence`"
- **Question specificity (Level 3):** the precise question detail, e.g., "`When did [A] start [B]?`"

The LLM generates alternative combinations along each axis for the remaining 7 leaves, producing distractor questions that share partial semantic overlap with q_i but differ in at least one axis. Each edge record is encoded as a nested JSON object {"k_i": {"`step`": "...", "`next`": "k_j"}} for internal edges, or {"k_i": {"`step`": "..."}} for leaf edges (no `next` field). Algorithm 2 formalizes the construction of a Compositional StateTree, and examples are provided in Appendix E.

The model must **(i)** navigate the tree using the same temporal discrimination as the Basic StateTree; **(ii)** accumulate the `step` fragments along the correct root-to-leaf path; and **(iii)** aggregate them into the final question before answering it based on the dialogue content.

3.2 Long-Context Multi-turn Dialogue Reinforcement Learning

We adopt GRPO [25] with curriculum RL training and a combined reward function.

Curriculum RL Training. Training directly on high-depth trees is difficult for LLMs and often leads to instability. We therefore adopt a four-stage curriculum that gradually increases task difficulty [37, 38, 39].

- **Warmup: Direct Dialogue QA.** We first warm up the model with RL on the original dialogue QA data, establishing basic long-context comprehension before introducing tree-structured tasks.
- **Stage 1: Basic StateTree (D=2).** We introduce the Basic StateTree at depth D=2 (4 leaves, 6 edge records scattered across sessions), requiring the model to perform cross-session retrieval and temporal discrimination as fundamental skills to reach the target leaf.
- **Stage 2: Basic StateTree (D=3).** We increase the depth to D=3 (8 leaves, 14 edge records), adding an additional hop that forces the model to strengthen multi-hop reasoning.
- **Stage 3: Compositional StateTree (D=3).** Stage 3 replaces the Basic StateTree with the Compositional StateTree, adding the compositional skill of accumulating step fragments along the correct path and aggregating them into the final question before answering.

Combined Reward. Training with a single reward source is suboptimal: exact-match (EM) reward is overly rigid and rejects semantically correct answers that differ in surface form, while LLM-as-a-Judge suffers from scoring inconsistency and higher computational overhead (see ablation in Table 8). To provide a stable and semantically faithful optimization signal, we assign the reward r_i via a combined function that integrates rule-based verification with LLM-as-a-Judge [33]. We explicitly require the model to output its final answer within `\boxed{...}` in the training prompt (Appendix D), ensuring the extraction of an unambiguous answer y_{ans}. Formally, given question x, extracted answer y_{ans}, and ground-truth answer y_{gold}, the reward is defined as:

$$r_\phi(x, y) = \max\left(r_{\text{EM}}(y_{\text{ans}} = y_{\text{gold}}),\ r_{\text{LLM}}(x, y_{\text{ans}}, y_{\text{gold}})\right) \tag{1}$$

where $r_{\text{EM}}(\cdot)$ is the indicator function enforcing exact string matching, and $r_{\text{LLM}}(\cdot) \in \{0, 1\}$ is a binary semantic equivalence score produced by a Qwen2.5-1.5B-Instruct [40]. The judge operates at temperature 0 with prompt templates (Appendix D.6.1) to guarantee deterministic outputs.

4 Experiments

4.1 Experimental Setup

Training Setup. We run experiments on Qwen2.5-7B-Instruct [40], Qwen2.5-14B-Instruct [40], and Qwen3-8B [41]. We use GRPO with group size 8, prompt batch size 64 for 7B and 8B models and 32 for 14B model, learning rate 1×10^{-6}, gradient clipping 1.0, and KL penalty β=0.001. Rollouts are sampled with temperature 0.6 and top-p=0.95, with a maximum output length of 4,096 tokens. The curriculum consists of four stages: warm-up (40 steps), Stage 1 (100 steps), Stage 2 (100 steps), and Stage 3 (60 steps). To prevent catastrophic forgetting of general reasoning capabilities, we mix 2,500 samples from the DAPO-Math dataset [42] into Stage 3. The warm-up stage uses the 616 training samples in the original direct QA format without trees; Stage 1–3 augment the same 616 samples with StateTree. The 770 test samples and 154 validation samples both retain the original direct QA format without any StateTree augmentation. All models are trained on 32×H20 GPUs.

Table 2: Main results on the LoCoMo benchmark across different task dimensions.

Model	Multi Hop			Temporal			Open Domain			Single Hop			Average		
	ACC	F1	BLEU	ACC	F1	BLEU	ACC	F1	BLEU	ACC	F1	BLEU	ACC	F1	BLEU
GPT-4o	68.84	**32.25**	**23.10**	14.47	10.25	10.45	30.00	18.47	16.80	**64.54**	**49.34**	**40.19**	**52.73**	**36.20**	**29.47**
QwenLong-L1-32B	64.49	29.53	22.27	34.59	**24.35**	**20.19**	48.00	**29.43**	**27.02**	44.68	30.37	26.20	46.36	28.92	24.31
R1-Distill-Qwen-32B	**70.29**	26.80	17.45	**33.33**	15.98	12.35	**52.00**	25.19	20.98	52.01	23.04	18.00	51.43	22.40	16.93
Qwen2.5-7B-Instruct	63.77	21.34	15.23	23.90	12.67	10.58	50.00	16.88	14.08	44.92	21.96	17.72	44.29	19.60	15.56
SEALONG-7B	63.77	27.50	20.65	20.75	15.16	11.72	48.00	25.80	23.53	47.99	32.01	28.38	45.19	27.32	23.24
LoongRL-7B	60.87	28.84	19.53	25.16	18.25	15.08	**56.00**	23.56	19.68	47.99	32.92	28.41	46.10	28.55	23.50
RL-MemAgent-7B	**64.49**	30.62	21.88	31.45	16.36	12.31	46.00	26.26	22.41	49.14	32.75	27.07	48.03	28.56	22.79
StateTree-7B	**64.49**	**31.20**	**25.38**	**32.70**	**33.97**	**29.36**	**56.00**	**36.11**	**32.31**	**50.35**	**36.00**	**32.18**	**49.61**	**34.73**	**30.39**
Qwen2.5-14B-Instruct	67.39	31.98	24.07	34.59	14.53	11.17	62.00	27.31	23.83	47.04	34.32	30.16	49.09	29.36	24.74
SEALONG-14B	69.57	35.64	25.65	32.70	15.85	12.95	58.00	26.37	22.08	47.28	32.73	28.76	48.96	29.36	24.50
LoongRL-14B	69.57	33.20	25.82	37.74	20.53	23.25	62.00	34.14	31.10	52.25	35.76	30.25	52.99	32.05	28.07
RL-MemAgent-14B	67.39	31.96	25.01	37.11	14.55	11.13	60.00	33.04	30.10	50.59	34.62	29.39	51.43	29.90	24.88
StateTree-14B	**75.36**	**38.40**	**27.90**	**41.51**	**40.72**	**33.56**	**66.00**	**39.65**	**37.18**	**62.88**	**38.37**	**33.26**	**60.91**	**38.94**	**32.62**
Qwen3-8B	64.49	31.32	22.13	29.56	18.91	15.31	48.00	21.4	18.55	43.03	26.37	23.02	43.64	25.39	20.98
StateTree-8B	**70.29**	**36.07**	**25.58**	**37.74**	**31.78**	**29.25**	**56.00**	**36.36**	**33.08**	**51.06**	**36.75**	**31.89**	**52.08**	**35.58**	**30.29**

Table 3: OOD generalization accuracy (%) on LongMemEval and PersonaMem-128k benchmarks.

Model	LongMemEval (128k)							PersonaMem (128k)							
	Temp.	Multi-Ses.	Know. Upd.	SS-User	SS-Asst.	SS-Pref.	Avg.	Latest Pref.	New Scen.	Align. Rec.	Shared Fact	Revisit Reas.	New Idea	Track Evol.	Avg.
QwenLong-L1-32B	**47.37**	**35.34**	**56.41**	**80.00**	**67.86**	**23.33**	**51.00**	**57.51**	**48.83**	**47.56**	**66.08**	**81.04**	**25.48**	**58.06**	**52.40**
R1-Distill-Qwen-32B	21.80	13.53	50.00	34.29	57.14	10.00	29.00	33.03	21.60	33.52	36.84	63.57	22.39	66.57	37.62
Qwen2.5-7B-Instruct	20.30	8.27	52.56	25.71	37.50	3.33	23.80	37.41	36.15	45.56	47.95	62.08	15.64	62.76	40.48
SEALONG-7B	28.57	28.57	64.10	75.71	76.79	16.67	45.40	37.99	39.44	46.99	47.37	64.68	16.80	63.64	41.66
LoongRL-7B	24.06	5.26	55.13	37.14	42.86	3.33	26.60	29.56	23.00	39.83	35.09	41.26	21.81	37.54	31.39
RL-MemAgent-7B	28.57	17.29	58.97	**78.57**	76.79	13.33	41.80	43.07	33.33	45.27	49.12	65.06	23.17	62.46	43.78
StateTree-7B	**29.32**	**31.58**	**67.95**	75.71	**76.79**	**23.33**	**47.40**	**46.42**	**39.91**	**51.00**	**49.71**	**68.03**	**23.36**	**69.21**	**47.30**
Qwen2.5-14B-Instruct	37.59	27.82	67.95	65.71	66.07	10.00	45.20	43.76	38.03	52.15	57.31	68.03	18.34	64.52	45.40
SEALONG-14B	40.60	30.08	67.95	72.86	66.07	13.33	47.80	45.84	38.97	51.29	61.40	68.77	15.44	66.28	46.02
LoongRL-14B	40.60	36.84	69.23	81.43	**89.29**	23.33	54.20	48.15	40.85	56.16	60.23	68.77	18.15	67.16	48.07
RL-MemAgent-14B	**42.11**	29.32	67.95	85.71	82.14	20.00	52.00	47.69	39.44	53.30	60.82	71.00	21.04	66.28	48.15
StateTree-14B	**42.11**	**48.12**	**70.51**	**87.14**	85.71	**36.67**	**59.00**	**51.62**	**46.48**	**59.89**	**64.33**	**77.32**	**23.55**	**70.09**	**52.59**
Qwen3-8B	35.34	39.10	66.67	72.86	69.64	16.67	49.20	53.81	39.44	37.82	48.54	69.52	16.60	58.36	45.36
StateTree-8B	**41.00**	**47.47**	**78.21**	**85.45**	**83.93**	**26.09**	**58.66**	**58.31**	**46.95**	**41.26**	**46.78**	**72.86**	**21.04**	**69.79**	**50.31**

Evaluation Benchmarks. We evaluate models across two dimensions. (i) Long-term conversation reasoning. We use LoCoMo [15] as the in-domain (ID) benchmark, and LongMemEval [16] and PersonaMem [27] as out-of-distribution (OOD) benchmarks to test generalization across unseen domains and extended contexts (10K–128K tokens). We follow official protocols: LoCoMo reports LLM-judged accuracy (evaluated by GPT-4o), token-level F1, and BLEU-1; LongMemEval reports LLM-judged accuracy (evaluated by GPT-4o); PersonaMem reports exact-match accuracy. (ii) General short-context reasoning. To verify that our long-context RL training does not induce catastrophic forgetting or degrade foundational reasoning, we additionally evaluate on standard short-context benchmarks: MMLU [43], MATH-500 [44], and IFEval [45]. All evaluations share a unified inference setup, i.e., temperature 0.6, with up to 128K input tokens and 4096 output tokens.

Baselines. We compare our StateTree-trained models against: **(i)** leading frontier models and long-context reasoning models, including GPT-4o [46], QwenLong-L1-32B [33], and R1-Distill-Qwen-32B [24]; **(ii)** memory-augmented models that enhance long-context reasoning, including the SFT-based method SEALONG [31] and RL-based methods LoongRL [35] and RL-MemAgent [47].

4.2 Main Results

StateTree improves long-term dialogue reasoning with high data efficiency. As shown in Table 2, StateTree delivers consistent performance gains over both SFT and standard RL baselines across parameter scales and base architectures on LoCoMo. StateTree-14B achieves 60.91% average accuracy, yielding a +11.82% absolute improvement over its base model (Qwen2.5-14B-Instruct), while maintaining stable relative gains at the 7B scale. Notably, despite training on only 616 examples, it matches or exceeds the absolute performance of substantially larger models (e.g., surpassing GPT-4o by 8.18% and QwenLong-L1-32B by 14.55%), underscoring its high sample efficiency. The most pronounced lift occurs in Temporal reasoning (F1 +26.19 for 14B), which directly stems from the tree structure's explicit timestamp comparison at each branching step.

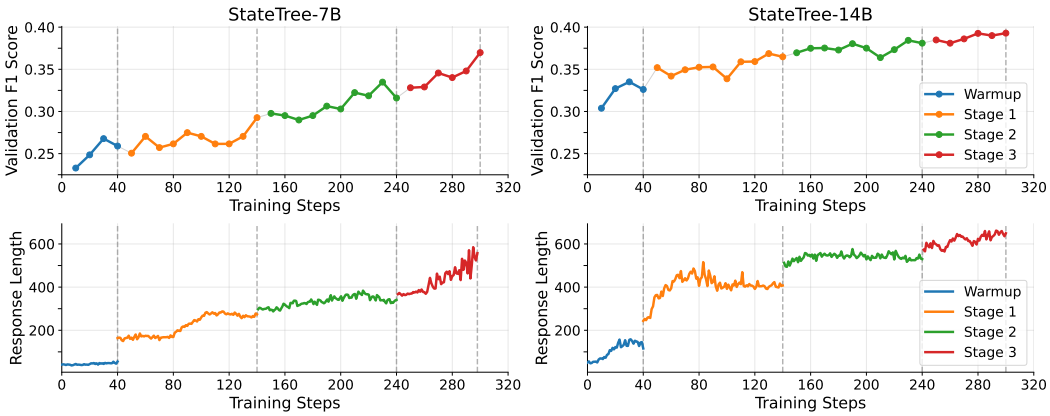


Figure 4: Reasoning F1 score and response lengths throughout RL training on the validation set.

Robust out-of-distribution and length generalization. To evaluate out-of-distribution generalization, we evaluate StateTree on LongMemEval [16] and PersonaMem-128k [27]. Trained exclusively on 10K-token contexts, StateTree generalizes to unseen benchmarks with up to 128K tokens. As shown in Table 3, StateTree-14B achieves 59.00% on LongMemEval and 52.59% on PersonaMem-128k, outperforming both base models and strong baselines including QwenLong-L1-32B and RL-MemAgent-14B. The widening performance gap at longer contexts (Table 4) demonstrates that the learned path-tracing strategy extrapolates effectively beyond the 10K training horizon, validating our curriculum design as an alternative to prohibitive full-length RL costs (ℂ3).

Table 4: Average accuracy (%) across context lengths.

Model	LoCoMo	PersonaMem		LongMemEval
	10K	32K	128K	128K
Qwen2.5-7B-Instruct	44.29	53.14	40.48	23.80
StateTree-7B	**49.61**+5.32	**61.29**+8.15	**47.30**+6.82	**47.40**+23.60
Qwen2.5-14B-Instruct	49.09	58.91	45.40	45.20
StateTree-14B	**60.91**+11.82	**64.52**+5.61	**52.59**+7.19	**59.00**+13.80
Qwen3-8B	43.64	58.23	45.36	49.20
StateTree-8B	**52.08**+8.44	**62.82**+4.59	**50.31**+4.95	**58.66**+9.46

StateTree preserves short-context abilities without capability trade-offs. A common concern with long-context RL is the degradation of general reasoning. As shown in Table 5, StateTree maintains competitive performance across standard short-context benchmarks (MMLU, MATH, IFEval). Notably, the 8B variant exhibits zero degradation, achieving a marginal +0.1% average gain over the base Qwen3-8B, while the 7B and 14B models incur minimal drops of only -0.2% and -0.6%, respectively.

Table 5: Short-context reasoning performance (%).

Model	MMLU	MATH	IFEval	Avg.
Qwen2.5-7B-Instruct	73.4	**76.0**	**71.2**	**73.5**
StateTree-7B	**73.5**+0.10	**76.0**+0.00	70.5-0.70	73.3-0.20
Qwen2.5-14B-Instruct	79.4	**83.4**	**81.0**	**81.3**
StateTree-14B	**79.8**+0.40	83.0-0.40	79.3-1.70	80.7-0.60
Qwen3-8B	76.9	**78.2**	85.0	80.0
StateTree-8B	**77.0**+0.10	78.1-0.10	**85.2**+0.20	**80.1**+0.10

4.3 Analysis

Multi-stage curriculum RL training sustains improvements. As shown in Figure 4, the average F1 score on the validation set rises monotonically across training stages, and the average response length steadily increases alongside it, indicating that the model learns to produce detailed reasoning traces as task complexity increases. The sustained growth in both metrics with no sign of reward hacking confirms that the combined reward provides a stable training signal and that the multi-stage curriculum effectively scales model reasoning.

Emergent reasoning behaviors align with task design. Qualitative analysis (Figure 1 in Section 1 and Appendix K) reveals four emergent behaviors that directly explain the quantitative gains: (i) *Cross-session retrieval.* Rather than stopping at the first matching session, StateTree enumerates all relevant records across sessions before aggregating a complete answer. (ii) *Multi-hop reasoning.* Rather than identifying isolated facts, StateTree chains multiple retrieval steps and synthesizes scattered evidence into a coherent inference. (iii) *Temporal reasoning.* Rather than relying on

superficial cues such as "recently," StateTree extracts concrete session timestamps and performs explicit chronological comparison at each fork. (iv) *Knowledge update.* Rather than conflating old and new mentions, StateTree correctly tracks the latest state by privileging information from more recent sessions. These behaviors arise from the StateTree pseudo-task design, confirming that our auxiliary training signal successfully imparts targeted reasoning patterns that can transfer beyond the training distribution.

4.4 Ablation Study

Every curriculum stage contributes. Table 6 ablates each stage individually, where the omitted stage is skipped and the remaining stages retain their original step allocations. Removing Stage 1 (D=2) causes the largest drop (-4.93), confirming that shallow tree navigation is the most critical foundational skill. Removing warm-up follows (-4.42), showing that grounding in basic conversation QA is essential before tree tasks. Removing Stage 2 (D=3) and Stage 3 (Compositional) degrade performance by -3.38 and -2.86, respectively, indicating their roles in strengthening multi-hop temporal reasoning and enabling compositional reasoning. Appendix H provides a per-category breakdown that further corroborates the curriculum design. This sustained uplift validates our staged design: by progressively introducing foundational comprehension, temporal discrimination, and compositional reasoning, the curriculum enables stable skill accumulation without catastrophic forgetting, effectively distilling transferable long-context primitives from a compact training set.

Table 6: Curriculum stage ablation.

Variant	LoCoMo	LongMemEval	PersonaMem
StateTree-7B (full)	**49.61**	**47.40**	**47.30**
w/o warm-up	45.19-4.42	43.00-4.40	44.00-3.30
w/o Stage 1 (D=2)	44.68-4.93	42.60-4.80	43.78-3.52
w/o Stage 2 (D=3)	46.23-3.38	44.20-3.20	45.10-2.20
w/o Stage 3 (Compositional)	46.75-2.86	44.80-2.60	45.51-1.79

Task structure drives the training signal. To confirm that StateTree's efficacy stems from the synergistic integration of key-paired forks, cross-session scattering, and identifier opacity—rather than from merely inserting structured records or prompting longer reasoning, we design four variants to isolate the contribution of each structural design choice (Table 7).

Table 7: Temporal discrimination ablation.

Variant	LoCoMo	LongMemEval	PersonaMem
Qwen2.5-7B-Instruct	44.29	23.80	40.48
w/ CoT prompt	45.19+0.90	24.60+0.80	41.11+0.63
StateTree-7B	**49.61**	**47.40**	**47.30**
w/ Independent Random Keys	45.45-4.16	25.20-22.20	41.03-6.27
w/ Single Session	47.79-1.82	42.80-4.60	43.96-3.34
w/ Entity Keys	48.44-1.17	45.60-1.80	45.98-1.32

- *w/ CoT prompt*: While Figure 4 shows that StateTree training naturally elicits progressively longer reasoning traces, prompting the base model via a modified system prompt (Appendix D.5) that instructs it to "think step by step" yields only marginal gains. This confirms that generic CoT cannot substitute for explicit structural navigation. The core bottleneck lies in unstructured cross-session retrieval and temporal discrimination, not insufficient reasoning depth.
- *w/ Independent Random Keys*: assigning independent random keys to each fork's competing edge records eliminates shared identifiers, preventing the model from locating competing records or performing temporal comparison. The resulting performance demonstrates that StateTree's gains originate specifically from the shared-key mechanism enabling explicit fork resolution, rather than from merely injecting records or arbitrary syntactic noise during RL training.
- *w/ Single Session*: preserving key pairing but placing all records within one session eliminates cross-session retrieval. The intermediate performance confirms that cross-session distribution and temporal discrimination provide complementary, non-redundant learning signals.
- *w/ Entity Keys*: replacing random UUIDs with common words from a word pool (Appendix C.7) captures 80% of the full gain. The remaining deficit reveals that semantically interpretable keys invite shallow lexical shortcuts, whereas opaque identifiers force genuine temporal reasoning.

Combined reward is critical for stable optimization. Table 8 compares reward designs under identical training settings. The combined reward balances strict correctness with semantic flexibility. EM alone rejects valid paraphrases due to surface-form rigidity, while standalone LLM-Judge suffers from scoring inconsistency and higher

Table 8: Reward function ablation.

Reward	LoCoMo	LongMemEval	PersonaMem
Combined Reward	**49.61**	**47.40**	**47.30**
Exact Match only	47.79-1.82	46.80-0.60	46.68-0.62
LLM-as-a-Judge only	48.70-0.91	46.60-0.80	46.75-0.55
Token-level F1	47.14-2.47	45.00-2.40	45.29-2.01
Two-way Substr. EM	44.68-4.93	42.40-5.00	43.31-3.99
ROUGE-L	42.60-7.01	39.40-8.00	41.29-6.01

compute overhead. Dense lexical metrics (Token F1, Substring EM) and ROUGE-L underperform by rewarding incorrect partial matches or superficial n-gram overlap. Thus, the max-selection mechanism is essential to provide a dense yet reliable training signal aligned with answer correctness.

5 Conclusion

We present StateTree, a data-driven RL pseudo-task that reframes long-term dialogue reasoning as a temporally constrained path-search problem. By embedding tree-structured auxiliary tasks into authentic multi-session dialogues and optimizing them via a structured curriculum with combined-reward GRPO, StateTree elicits cross-session retrieval, temporal reasoning, knowledge update, and compositional multi-hop reasoning from only 616 examples at 10K context. Empirically, it yields substantial gains across both in-domain and out-of-domain benchmarks, even surpassing larger baselines. Crucially, the trained models generalize from 10K to 128K contexts without degrading short-context capabilities and exhibit emergent reasoning behaviors aligned with the StateTree design. StateTree establishes a scalable, data-efficient paradigm for long-term dialogue reasoning.

Acknowledgments and Disclosure of Funding

This work was partly supported by the NSFC under No. 62402418, the "Pioneer and Leading Goose" R&D Program of Zhejiang under No. 2026C02A1233 and 2025C02034, the Key R&D Program of Ningbo under No. 2024Z115, and the Ningbo Yongjiang Talent Project.

References

[1] Sheshera Mysore, Zhuoran Lu, Mengting Wan, Longqi Yang, Bahareh Sarrafzadeh, Steve Menezes, Tina Baghaee, Emmanuel Barajas Gonzalez, Jennifer Neville, and Tara Safavi. Pearl: Personalizing large language model writing assistants with generation-calibrated retrievers. In Sachin Kumar, Vidhisha Balachandran, Chan Young Park, Weijia Shi, Shirley Anugrah Hayati, Yulia Tsvetkov, Noah Smith, Hannaneh Hajishirzi, Dongyeop Kang, and David Jurgens, editors, *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 198–219, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[2] Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. Are large language models capable of generating human-level narratives? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17659–17681, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[3] Wenyue Hua, Lei Li, Shuyuan Xu, Li Chen, and Yongfeng Zhang. Tutorial on large language models for recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1281–1283, 2023.

[4] Yuyuan Li, Yizhao Zhang, Weiming Liu, Xiaohua Feng, Zhongxuan Han, Chaochao Chen, and Chenggang Yan. Multi-objective unlearning in recommender systems via preference guided pareto exploration. *IEEE Transactions on Services Computing*, 2025.

[5] Yuyuan Li, Xiaohua Feng, Chaochao Chen, and Qiang Yang. A Survey on Recommendation Unlearning: Fundamentals, Taxonomy, Evaluation, and Open Questions . *IEEE Transactions on Knowledge & Data Engineering*, (01):1–20, 2025.

[6] Chaochao Chen, Yizhao Zhang, Yuyuan Li, Jun Wang, Lianyong Qi, Xiaolong Xu, Xiaolin Zheng, and Jianwei Yin. Post-training attribute unlearning in recommender systems. *ACM Transactions on Information Systems*, 43(1):1–28, 2024.

[7] Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. Travelplanner: a benchmark for real-world planning with language agents. In *Proceedings of the 41st International Conference on Machine Learning*, pages 54590–54613, 2024.

[8] Yiming Du, Bingbing Wang, Yang He, Bin Liang, Baojun Wang, Zhongyang Li, Lin Gui, Jeff Z Pan, Ruifeng Xu, and Kam-Fai Wong. Bridging the long-term gap: A memory-active policy for multi-session task-oriented dialogue. *arXiv e-prints*, pages arXiv–2505, 2025.

[9] 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, pages 19724–19731, 2024.

[10] Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. In *The Fourteenth International Conference on Learning Representations*, 2026.

[11] Zhan Ling, Kang Liu, Kai Yan, Yifan Yang, Weijian Lin, Ting-Han Fan, Lingfeng Shen, Zhengyin Du, and Jiecao Chen. Longreason: A synthetic long-context reasoning benchmark via context expansion. *arXiv preprint arXiv:2501.15089*, 2025.

[12] Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, Huajun Chen, and Ningyu Zhang. Lightmem: Lightweight and efficient memory-augmented generation. In *The Fourteenth International Conference on Learning Representations*, 2026.

[13] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380, 2018.

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

[15] 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)*, pages 13851–13870, 2024.

[16] 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.

[17] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[18] Aonian Li, Bangwei Gong, Bo Yang, Boji Shan, Chang Liu, Cheng Zhu, Chunhao Zhang, Congchao Guo, Da Chen, Dong Li, et al. Minimax-01: Scaling foundation models with lightning attention. *arXiv preprint arXiv:2501.08313*, 2025.

[19] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*, 2023.

[20] Ning Shang, Li Lyna Zhang, Siyuan Wang, Gaokai Zhang, Gilsinia Lopez, Fan Yang, Weizhu Chen, and Mao Yang. Longrope2: Near-lossless llm context window scaling. *arXiv preprint arXiv:2502.20082*, 2025.

[21] Akshara Prabhakar, Zuxin Liu, Ming Zhu, Jianguo Zhang, Tulika Awalgaonkar, Shiyu Wang, Zhiwei Liu, Haolin Chen, Thai Hoang, Juan Carlos Niebles, et al. Apigen-mt: Agentic pipeline for multi-turn data generation via simulated agent-human interplay. *arXiv preprint arXiv:2504.03601*, 2025.

[22] Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models. *arXiv preprint arXiv:2505.24864*, 2025.

[23] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025.

[24] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[25] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[26] 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.

[27] Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J Taylor, and Dan Roth. Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale. *arXiv preprint arXiv:2504.14225*, 2025.

[28] Naen Xu, Hengyu An, Shuo Shi, Jinghuai Zhang, Chunyi Zhou, Changjiang Li, Tianyu Du, Zhihui Fu, Jun Wang, and Shouling Ji. When agents "misremember" collectively: Exploring the mandela effect in LLM-based multi-agent systems. In *The Fourteenth International Conference on Learning Representations*, 2026.

[29] Yebo Wu, Feng Liu, Ziwei Xie, Changwang Zhang, Jun Wang, and Li Li. Tsembed: Unlocking task scaling in universal multimodal embeddings. In *European Conference on Computer Vision*, pages 1–19. Springer, 2026.

[30] Yebo Wu, Han Jin, Zhijiang Guo, and Li Li. Spotlight and shadow: Attention-guided dual-anchor introspective decoding for mllm hallucination mitigation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 13219–13233, 2026.

[31] Siheng Li, Cheng Yang, Zesen Cheng, Lemao Liu, Mo Yu, Yujiu Yang, and Wai Lam. Large language models can self-improve in long-context reasoning. *arXiv preprint arXiv:2411.08147*, 2024.

[32] Yanyang Li, Shuo Liang, Michael Lyu, and Liwei Wang. Making long-context language models better multi-hop reasoners. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2462–2475, 2024.

[33] Fanqi Wan, Weizhou Shen, Shengyi Liao, Yingcheng Shi, Chenliang Li, Ziyi Yang, Ji Zhang, Fei Huang, Jingren Zhou, and Ming Yan. Qwenlong-l1: Towards long-context large reasoning models with reinforcement learning. *arXiv preprint arXiv:2505.17667*, 2025.

[34] Greg Kamradt. Needle in a haystack - pressure testing llms, 2023.

[35] Siyuan Wang, Gaokai Zhang, Li Lyna Zhang, Ning Shang, Fan Yang, Dongyao Chen, and Mao Yang. LoongRL: Reinforcement learning for advanced reasoning over long contexts. In *The Fourteenth International Conference on Learning Representations*, 2026.

[36] Sarah Dean and Jamie Morgenstern. Preference dynamics under personalized recommendations. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 795–816, 2022.

[37] Yebo Wu, Jingguang Li, Zhijiang Guo, and Li Li. Developmental federated tuning: A cognitive-inspired paradigm for efficient LLM adaptation. In *The Fourteenth International Conference on Learning Representations*, 2026.

[38] Yebo Wu, Jingguang Li, Zhijiang Guo, and Li Li. Don't reinvent the wheel, just realign the spokes: Resource-efficient federated fine-tuning via rank-wise expert assembly. In Tong Zhang, Miroslav Dudik, Martin Jaggi, Alekh Agarwal, Sharon Li, Dale Schuurmans, Jerry Zhu, Felix Berkenkamp, Hanze Dong, and Alberto Bietti, editors, *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 136193–136214. PMLR, 06–11 Jul 2026.

[39] Yebo Wu, Jingguang Li, Chunlin Tian, KaHou Tam, Zhijiang Guo, and Li Li. Beyond end-to-end: Dynamic chain optimization for private LLM adaptation on the edge. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens, editors, *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18419–18435, San Diego, California, United States, July 2026. Association for Computational Linguistics.

[40] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[41] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[42] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

[43] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[44] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

[45] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

[46] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[47] 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 *The Fourteenth International Conference on Learning Representations*, 2026.

[48] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

A Limitations and Broader Impacts

A.1 Limitations

Our work has several limitations. First, all training data are derived from a single English dialogue dataset (LoCoMo), so the generalization of StateTree to other languages or dialogue domains remains to be validated. Second, the Compositional StateTree relies on an LLM (GPT-4o) to decompose questions and generate distractors; while we enforce structural validation, the quality of decomposition is bounded by the LLM's capability and may introduce subtle semantic biases. Finally, we use fixed binary-tree depths (D=2 and D=3) throughout training; adaptive depths or dynamic tree structures could further improve efficiency but are not explored here.

A.2 Broader Impacts

StateTree aims to improve the reliability of personalized assistants in long-term interactions, which can enhance user experience in education, healthcare, and personal productivity. However, more powerful reasoning capabilities could also be misused to extract sensitive information from conversation histories or enable unauthorized surveillance. We encourage developers to deploy such systems with appropriate privacy safeguards, user consent mechanisms, and data retention policies.

B Licenses and Asset Usage

B.1 Existing Assets

We use the publicly available LoCoMo [15], LongMemEval [16], and PersonaMem [27] datasets. Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct, and Qwen2.5-1.5B-Instruct are distributed under the Apache License 2.0, as is Qwen3-8B [41]. Qwen2.5-1.5B-Instruct serves as the LLM judge in our combined reward function. GPT-4o is accessed through the OpenAI API.

B.2 LLM Usage Declaration

LLMs are used as non-standard components in two parts of our methodology. (i) **Data construction:** We use GPT-4o (temperature 0.7) to decompose target questions into semantic axes and generate distractor steps for the Compositional StateTree (Section 3.1 and Appendix C.4.1). (ii) **Reward evaluation:** We use Qwen2.5-1.5B-Instruct (temperature 0) as an LLM-as-a-Judge evaluator within the combined reward function (Section 3.2).

C Dataset Construction Details

C.1 Source Data

All training data are derived from the LoCoMo dataset [15], which contains multi-session dialogues between two speakers. Each conversation comprises approximately 13 sessions with interleaved `DATE:` markers indicating session timestamps. We use the same 616 conversation–question pairs across all four curriculum stages; only the *task formulation* changes between stages.

C.2 Dataset Split

We filter out adversarial and empty-answer questions, then split the remaining pairs 50/50. The first half uses an 80/20 train/validation split (616 and 154 samples), and the second half forms the held-out test set.

The validation and test sets use the original direct QA format (identical to the Warm-up stage) without any tree augmentation or record insertion. This ensures that evaluation measures the model's end-task performance on clean conversation data. The same training set of 616 samples is reused across all four curriculum stages; only the task formulation (tree structure and prompt) changes between stages.

C.3 Dataset Statistics and Construction Parameters

Table 9 summarizes the dataset statistics and data construction parameters for each curriculum stage.

Table 9: Dataset statistics and data construction parameters for each curriculum stage. "UUID format" refers to the identifier format used for tree node keys. "Insert mode" specifies where records are placed within the dialogue text.

Parameter	**Warm-up**	**Stage 1**	**Stage 2**	**Stage 3**
Tree type	None (direct QA)	Basic StateTree	Basic StateTree	Compositional StateTree
Tree depth D	–	2	3	3
Leaves	–	4	8	8
Edge records	0	6	14	14
Samples	616	616	616	616
Avg. length (tokens)	∼10K	∼10K	∼10K	∼10K
UUID format	–	Standard UUID4	Standard UUID4	Standard UUID4
Record format	–	`{UUID: VALUE}`	`{UUID: VALUE}`	`{UUID: {step, next}}`
Insert mode	–	Inside speaker quotes	Inside speaker quotes	Inside speaker quotes
Temporal discrimination	–	Yes (session DATE)	Yes (session DATE)	Yes (session DATE)
Distractor pool size	–	64	64	8 (from decomposition)

C.4 Question Decomposition

For Stage 3 (Compositional StateTree), each target question is decomposed into a $2 \times 2 \times 2$ hierarchy using an LLM. Given a target question such as "When did Jon start expanding his studio's social media presence?", the decomposition produces three step fragments along semantic axes:

- **Step 1** (Person selection): `[A] Jon`
- **Step 2** (Event category): `[B] expanding his studio's social media presence`
- **Step 3** (Question detail): `When did [A] start [B]?`

The full decomposition tree for one conversation contains 8 such paths (one per leaf), each branching on person at level 1, event category at level 2, and question specificity at level 3. Table 10 shows the structure of a complete decomposition tree.

Table 10: Semantic decomposition example for the target question "*When did Jon start expanding his studio's social media presence?*" (answer: "April, 2023"). ★ marks the target leaf. The correct path aggregates: step_1 = "[A] Jon", step_2 = "[B] expanding his studio's social media presence", step_3 = "When did [A] start [B]?", yielding the full question after substitution.

Level 1	**Level 2 (Event)**	**Level 3 (Question)**
[A] Jon	[B] expanding his studio's social media presence	★ *When did [A] start [B]?* *What did [A] do while [B]?*
	[B] going to a fair for exposure	*When did [A] start [B]?* *What events did [A] participate in while [B]?*
[A] Gina	[B] launching an ad campaign	*When did [A] start [B]?* *Why did [A] decide to start [B]?*
	[B] teaming up with a local artist	*When did [A] start [B]?* *What did [A] do while [B]?*

C.4.1 Decomposition LLM Prompt

To generate the $2 \times 2 \times 2$ decomposition trees for Stage 3, we use GPT-4o with temperature 0.7. Each target question is processed with the following system and user prompts. The system prompt specifies the tree structure, output schema, and quality constraints; the user prompt provides the target question along with the full question pool for that conversation.

Decomposition System Prompt (Stage 3 Data Construction)

You are a dataset construction assistant. Your task is to decompose a target question into a $2 \times 2 \times 2$ binary discrimination tree for a reading comprehension benchmark.
The tree has 3 levels:

- Level 1 (root $\rightarrow$ 2 branches): Person selection. Each branch carries step_1 = "[A] PersonName"
- Level 2 (each person $\rightarrow$ 2 branches): Event category. Each branch carries step_2 = "[B] gerund_phrase". The gerund phrase must be concrete and specific (e.g., "attending the LGBTQ support group"), NOT a vague category (e.g., "community activities").
- Level 3 (each event $\rightarrow$ 2 branches): Question specificity. Each branch carries step_3 = a question template that MUST contain BOTH [A] and [B] as placeholders. The two questions under the same event MUST ask about different factual aspects.

CRITICAL: step_3 must ask about CONCRETE FACTS that have definite answers — things like time, place, people involved, specific actions taken, specific results/outcomes, or specific objects.
GOOD examples (factual): "When did [A] [B]?", "What did [A] make while [B]?", "Where did [A] go for [B]?"
BAD examples (subjective/generic): "How did [A] feel about [B]?", "Why did [A] enjoy [B]?"
Exactly ONE of the 8 leaves must be marked is_target=true — the leaf whose step_3, after substituting [A] and [B], yields (or is semantically equivalent to) the original target question.
Rules:

1. The target person MUST appear first in the tree.
2. The target event MUST appear first under the target person.
3. The target leaf MUST be the first leaf under the target event.
4. Each event_gerund should be a present participle phrase derived from an actual question in the pool.
5. Every leaf's "answer" field MUST be copied verbatim from a question in the pool.
6. The two leaves under the same event MUST ask about genuinely DIFFERENT ASPECTS of the event.
7. Events for the same person should be distinct and both grounded in the question pool.
8. Events for the other person should also correspond to real pool questions but thematically contrast with the target person's events.
9. All answers must be short (a few words or a short phrase), copied from the pool.
10. EVERY step_3 MUST contain both [A] and [B] placeholders.

Decomposition User Prompt Template (Stage 3 Data Construction)

Target question (row {row_index}):
Q: {target_question}
sample_id: {sample_id}
Speakers in this conversation:
Speaker A: {speaker_a}
Speaker B: {speaker_b}
Below is the full question pool for this conversation ({num_questions} questions). Use these to find suitable distractor questions and events. Pick events and questions that are grounded in the actual conversation content.
— Question Pool —
{question_pool}
— End of Question Pool —
Please generate the $2 \times 2 \times 2$ decomposition tree JSON for this target question. Remember:

- step_2 must describe the EVENT/ACTIVITY itself, NOT reveal or hint at the answer
- EVERY step_3 MUST contain BOTH [A] and [B] placeholders
- step_3 must ask about CONCRETE FACTS
- The target person and target event come first in the tree
- Exactly 1 leaf is is_target=true
- Output raw JSON only, no markdown fencing

The question pool for each conversation is constructed by collecting all non-adversarial QA pairs associated with that conversation in the LoCoMo dataset, including their answers and evidence session numbers. Each generated decomposition is validated against a set of structural constraints (correct number of persons, events, leaves; presence of [A]/[B] placeholders; exactly one target leaf; semantic diversity of sibling step_3 values) and regenerated if validation fails.

C.5 Distractor Question Pool

At each StateTree stage (1–3), non-target leaves are populated with *distractor questions* from the same conversation and training split (excluding the target). Their semantic relevance to the dialogue makes them strong distractors. For Stage 3, distractors are generated by the $2 \times 2 \times 2$ decomposition (Appendix C.4), with each non-target leaf differing in at least one semantic axis.

C.6 Record Insertion Strategy

Tree edge records must be embedded naturally within the conversation text so that they blend into the dialogue flow. We employ the following insertion strategy:

Insertion positions. For all tree-augmented stages (Stages 1–3), records are inserted *inside speaker quotations*: within lines of the form `Speaker said, "..."`, records are placed at sentence boundaries (after periods, exclamation marks, or question marks) inside the quoted text. This prevents artifacts such as a JSON record appearing immediately before a `DATE:` line or between speaker turns.

Balanced session distribution. Records are distributed across conversation sessions using a least-loaded-first allocation strategy. For each record to be inserted, we:

1. Group available insertion slots by their enclosing session.
2. Identify the session(s) with the fewest records already assigned.
3. Randomly select a slot from the least-loaded session(s).

Same-key separation constraint. When the same UUID key appears in two records (as required by the temporal discrimination mechanism), these two records are guaranteed to be inserted into *different* sessions. This is enforced by tracking which sessions have already been used for each key and excluding them from the candidate set.

Post-insertion temporal ordering. After all records are inserted, a verification pass checks each temporal discrimination fork: the correct-path record must reside in a strictly more recent session (by parsed `DATE:` timestamp) than its distractor counterpart. If any violation is detected, the two records' positions in the context are swapped. This swap procedure ensures the temporal ordering invariant holds regardless of the initial random placement.

C.7 Entity Key Pool for Ablation

In the *w/ Entity Keys* ablation (Section 4.4), each UUID is replaced with a high-frequency concrete noun sampled without replacement from the 14-word pool below. This ensures that performance differences arise from semantic key content rather than topical relevance.

Table 11: Entity key pool used in the *w/ Entity Keys* ablation.

garden	mirror	temple	anchor	lantern	ribbon	candle
shield	bridge	castle	meadow	feather	saddle	trumpet

For example, under $D{=}3$ (7 keys required), a random draw might produce the edge records:

```
{‘‘garden’’: ‘‘bridge’’}, {‘‘garden’’: ‘‘meadow’’}, {‘‘bridge’’: ‘‘castle’’}, ...
```

instead of the standard UUID format `{‘‘f391e945-...’’: ‘‘8ada7089-...’’}`. The key pairing mechanism is preserved, but the model can now rely on the semantic familiarity of the word rather than performing pure string matching.

D Full Prompt Templates

D.1 System Prompt

The following system prompt is used across all curriculum stages. It instructs the model to produce a step-by-step reasoning trace enclosed in `<think>...</think>` tags and place the final answer inside `\boxed{...}`.

System Prompt

A conversation between User and Assistant. The User asks a question, and the Assistant solves it. The Assistant first thinks about the reasoning process in the mind and then provides the User with the answer. The reasoning process is enclosed within <think> </think> and answer is enclosed within \boxed{} tags, respectively, i.e., <think> reasoning process here </think> \boxed{answer here}.

D.2 Warm-up: Direct Dialogue QA

The Warm-up prompt is identical to the official LoCoMo evaluation prompt [15].

Warm-up Prompt

Based on the above conversations, write a short answer for the following question in a few words. Do not write complete and lengthy sentences. Answer with exact words from the conversations whenever possible.
Question: `{question}`

D.3 Stage 1 & 2: Basic StateTree (D=2 / D=3)

Stages 1 and 2 share the same prompt template; only the tree depth differs.

Basic StateTree Prompt (Stages 1 & 2)

In the conversation above, there are JSON records like `{''KEY'': ''VALUE''}` scattered throughout the dialogue text. They form UUID chains: starting from a given key, each value either points to the next key (a UUID) or contains the final question to answer.
Your task: start from key `''{root_uuid}''` and follow the chain to find the question, then answer it.
How to follow the chain:

1. Search the entire conversation for all records whose key matches the current UUID.
2. Each record sits inside a specific session. Look at the nearest preceding `''DATE: ...''` line to determine that record's time.
3. If the same key appears in multiple sessions, choose which record to use: among the remaining records, pick the one from the most recent session DATE.
4. Read the chosen value:
   - If it is a UUID, treat it as the next key and go back to step 1.
   - If it is a natural-language question, that is the question you must answer.
5. Once you find the question, answer it based on the conversation content.

D.4 Stage 3: Compositional StateTree (D=3)

Compositional StateTree Prompt (Stage 3)

In the conversation above, there are JSON records like `{''UUID'': {''step'': ''...'', ''next'': ''UUID''}}` scattered throughout the dialogue text. They form a decision tree.
Each record maps a UUID key to a JSON object with:
• `''step''`: a piece of information to accumulate along the path
• `''next''`: the UUID of the next node to follow (absent at leaf nodes)
Your task: start from key `''{root_uuid}''` and follow the tree to a leaf, accumulating the step information at each edge. Then answer the question.
How to follow the tree:

1. Search the entire conversation for all records whose key matches the current UUID.
2. Each record sits inside a specific session. Look at the nearest preceding `''DATE: ...''` line to determine that record's session time.
3. If the same key appears in multiple sessions, pick the record from the most recent session DATE.
4. Read the `''step''` field and remember it.
5. If a `''next''` field exists, use it as the new key and go back to step 1.
6. If no `''next''` field exists, you have reached a leaf.

Once you reach a leaf, aggregate all the step information you collected along the path to form the final question. Then answer that question step by step based on the conversation content.

D.5 CoT Prompt Baseline

For the *w/ CoT prompt* ablation variant (Section 4.4), we modify the System Prompt to explicitly encourage step-by-step reasoning, without any StateTree structure or RL training. This baseline tests whether generic reasoning prompts can replicate the gains from explicit structural navigation.

CoT System Prompt

A conversation between User and Assistant. The User asks a question, and the Assistant solves it. The Assistant first thinks about the reasoning process **step by step** in the mind, carefully analyzing the question and searching through the conversation for relevant information. Then the Assistant provides the User with the answer. The reasoning process is enclosed within <think> </think> and answer is enclosed within \boxed{ } tags, respectively, i.e., <think> reasoning process here </think> \boxed{answer here}.

D.6 LLM Evaluation Prompts

D.6.1 LoCoMo Evaluation Prompt

LoCoMo LLM Judge Prompt

Your task is to label an answer to a question as "CORRECT" or "WRONG".
You will be given the following data: (1) a question (posed by one user to another user), (2) a 'gold' (ground truth) answer, (3) a generated answer which you will score as CORRECT/WRONG.
The point of the question is to ask about something one user should know about the other user based on their prior conversations. The gold answer will usually be a concise and short answer that includes the referenced topic, for example:
Question: Do you remember what I got the last time I went to Hawaii?
Gold answer: A shell necklace
The generated answer might be much longer, but you should be generous with your grading — as long as it touches on the same topic as the gold answer, it should be counted as CORRECT.
For time related questions, the gold answer will be a specific date, month, year, etc. The generated answer might be much longer or use relative time references (like 'last Tuesday' or 'next month'), but you should be generous with your grading — as long as it refers to the same date or time period as the gold answer, it should be counted as CORRECT. Even if the format differs (e.g., 'May 7th' vs '7 May'), consider it CORRECT if it's the same date.
Now it's time for the real question:
Question: `{question}`
Gold answer: `{gold_answer}`
Generated answer: `{generated_answer}`
Return the label CORRECT or WRONG in a json format with the key as "label". Do NOT include both CORRECT and WRONG in your response, or it will break the evaluation script.

D.6.2 LongMemEval Evaluation Prompts

LongMemEval contains multiple task categories, each with a dedicated judge prompt.

LongMemEval Standard Tasks Prompt

I will give you a question, a correct answer, and a response from a model. Please answer yes if the response contains the correct answer. Otherwise, answer no. If the response is equivalent to the correct answer or contains all the intermediate steps to get the correct answer, you should also answer yes. If the response only contains a subset of the information required by the answer, answer no.
Question: `{question}`
Correct Answer: `{answer}`
Model Response: `{response}`
Is the model response correct? Answer yes or no only.

LongMemEval Temporal Reasoning Prompt

I will give you a question, a correct answer, and a response from a model. Please answer yes if the response contains the correct answer. Otherwise, answer no. If the response is equivalent to the correct answer or contains all the intermediate steps to get the correct answer, you should also answer yes. If the response only contains a subset of the information required by the answer, answer no. In addition, do not penalize off-by-one errors for the number of days. If the question asks for the number of days/weeks/months, etc., and the model makes off-by-one errors (e.g., predicting 19 days when the answer is 18), the model's response is still correct.
Question: `{question}`
Correct Answer: `{answer}`
Model Response: `{response}`
Is the model response correct? Answer yes or no only.

LongMemEval Knowledge Update Prompt

I will give you a question, a correct answer, and a response from a model. Please answer yes if the response contains the correct answer. Otherwise, answer no. If the response contains some previous information along with an updated answer, the response should be considered as correct as long as the updated answer is the required answer.
Question: `{question}`
Correct Answer: `{answer}`
Model Response: `{response}`
Is the model response correct? Answer yes or no only.

LongMemEval Single-session Preference Prompt

I will give you a question, a rubric for desired personalized response, and a response from a model. Please answer yes if the response satisfies the desired response. Otherwise, answer no. The model does not need to reflect all the points in the rubric. The response is correct as long as it recalls and utilizes the user's personal information correctly.
Question: `{question}`
Rubric: `{answer}`
Model Response: `{response}`
Is the model response correct? Answer yes or no only.

E Examples of Inserted Records in Context

This section illustrates how tree records are embedded within conversation text at each stage.

E.1 Basic StateTree Record (Stages 1 & 2)

E.1.1 Case Study

In the Basic StateTree, each record is a flat key-value pair `{UUID: VALUE}` inserted at a natural dialogue boundary. The value is either another UUID (internal node) or a natural-language question (leaf node).

Basic StateTree: Record Inserted in Dialogue

Caroline said, "Thanks, Melanie!
`{"95c49518-3f28-45f4-85be-de9a66ea5fd9": "What are Caroline's plans for the summer?"}`.
I really appreciate it. Excited for the future! Bye!"

In this example, the UUID key `95c49518...` maps directly to a question string, indicating a leaf node. If this were an internal node, the value would be another UUID pointing to the next hop.

The example below illustrates a concrete instance with D=2.

Green marks the correct path to leaf (target question); red marks distractor path to distractor leaves.

Example of StateTree-augmented dialogue input (Basic D=2)

Session 1 — **2:24 pm, 14 Aug, 2023**
Caroline: That pic is cool! Representing inclusivity and diversity in my art is important to me. `{"113464e2-8d54-493c-874a-ebb3c07b9d48": "5ab90ad1-5990-4820-9aec-6525e8e2867e"}` I also use it to speak up for the community and push for acceptance. Here's a recent painting! ...
Session 2 — **1:56 pm, 8 May, 2023**
Melanie: So cool you found such a helpful group. What's it done for you? `{"113464e2-8d54-493c-874a-ebb3c07b9d48": "9320b92a-cb63-4ed7-b22e-d3521da21950"}`
...
Session 3 — **8:18 pm, 6 July, 2023**
Caroline: Glad you have support, Caroline! `{"5ab90ad1-5990-4820-9aec-6525e8e2867e": "When did Melanie paint a sunrise?"}` How's it going? ...
Session 4 — **7:55 pm, 9 June, 2023**
Melanie: Glad you have support, Caroline! `{"5ab90ad1-5990-4820-9aec-6525e8e2867e": "What pets does Melanie have?"}` Unconditional love ...
Session 5 — **1:50 pm, 17 Aug, 2023**

Caroline: Sounds great, Mel! `{"9320b92a-cb63-4ed7-b22e-d3521da21950": "What kind of pot did Mel and her kids make with clay?"}` We'll . . .

Session 6 — **1:36 pm, 3 July, 2023**

Melanie: Wow, Caroline! That's great! I just signed up for a pottery class yesterday. It's like therapy for me, letting me express myself and get creative. `{"9320b92a-cb63-4ed7-b22e-d3521da21950": "Would Melanie go on another roadtrip soon?"}` Have you . . .

Instruction: In the dialogue above, a correct question is hidden within key:value records encoded with UUID strings. Starting from key 113464e2-8d54-493c-874a-ebb3c07b9d48, locate each record, resolve forks by selecting the edge record in the most recent session, follow the path to a leaf, then answer the recovered question.

Walkthrough. Starting from root key **113464e2-8d54-493c-874a-ebb3c07b9d48**, the model finds two records in Session 1 (Aug 14) and Session 2 (May 8). Session 1 is more recent, so it follows **5ab90ad1-5990-4820-9aec-6525e8e2867e**. Next, it finds records for **5ab90ad1-5990-4820-9aec-6525e8e2867e** in Session 3 (July 6) and Session 4 (June 9). Session 3 is more recent, leading to the leaf "When did Melanie paint a sunrise?". The model answers this question from the dialogue content.

E.2 Compositional StateTree Record (Stage 3)

In the Compositional StateTree, each record is a nested JSON object carrying a `step` field (the semantic fragment) and optionally a `next` field (the UUID of the child node).

Compositional StateTree: Records Inserted in Dialogue

(Root node - Level 1, person selection)
. . . cherishing those happy moments and clinging to them is key."
`{''84858d3a-1f30-41cd-9882-afe161ca4970'': {''step'': ''[A] Caroline'', ''next'': ''8ada7089-43a6-404c-b249-3d778e012633''}}`.
Melanie said, "Yeah, same here Caroline."
(Internal node - Level 2, event selection)
. . . "I've been thinking about volunteering more at the community center."
`{''8ada7089-43a6-404c-b249-3d778e012633'': {''step'': ''[B] expanding his studio's social media presence'', ''next'': ''6916ce1a-c2d0-4146-9e8d-185110a68332''}}`.
Caroline said, "That sounds like a great idea!"
(Leaf node - Level 3, no "next" field)
. . . "Reminds me it's important to cultivate a loving and accepting environment."
`{''6916ce1a-c2d0-4146-9e8d-185110a68332'': {''step'': ''When did [A] start [B]?''}}`.
and shared a photo of a group of people. . .

The root record provides `step: "[A] Caroline"` (person selection), the intermediate record provides `step: "[B] expanding his studio's social media presence"` (event selection), and the leaf record provides `step: "When did [A] start [B]?"` (question detail) with no `next` field. By following the correct path and aggregating all steps, the model reconstructs the full question.

F Hyperparameters and Training Configuration

F.1 Group Relative Policy Optimization (GRPO)

We adopt GRPO to train our model. For each question q, its long context $\mathcal{L}$ (the StateTree-augmented dialogue $\mathcal{S}'$ defined in Section 3.1), and its ground-truth answer a from a dataset $\mathcal{D}$, GRPO samples a group of rollout trajectories $\{o_1, o_2, \cdots, o_G\}$ from the old policy $\pi_{\theta_{old}}$. The policy π_θ is then optimized by maximizing:

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{(\mathcal{L},q,a)\sim\mathcal{D},\,\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|\mathcal{L},q)}$$
$$\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\Big(\min\left[\rho_{i,t}(\theta)A_i, \text{clip}(\rho_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon)A_i\right] - \beta D_{\text{KL}}\big(\pi_\theta(\cdot|q, o_{i,<t})\|\pi_{\text{ref}}(\cdot|q, o_{i,<t})\big)\Big)\right] \tag{2}$$

where $\rho_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q,o_{i,<t})}$. Hyperparameters ε and β control the clipping range of the importance sampling ratio and the weight of the KL penalty term, respectively. The advantage A_i is computed by normalizing the trajectory rewards $\{r_1, r_2, \ldots, r_G\}$ for each rollout trajectory:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \cdots, r_G\})}{\text{std}(\{r_1, r_2, \cdots, r_G\})} \tag{3}$$

Here, r_i is the reward for trajectory o_i. Following [24, 48], we use a combined reward function to assign r_i and mitigate reward hacking.

Training hyperparameters are provided in Section 3.2 of the main text. Each experiment is run three times with different random seeds, and we report the average results across runs.

F.2 Compute Resources

All training runs are performed on an internal cluster with 32×NVIDIA H20 GPUs (96 GB). The training times are approximately 90 hours for Qwen2.5-7B-Instruct, 100 hours for Qwen3-8B, and 160 hours for Qwen2.5-14B-Instruct.

G Evaluation Details

G.1 Benchmark Descriptions

- **LoCoMo** [15]: A multi-session dialogue benchmark with four reasoning categories: Multi Hop, Temporal, Open Domain, and Single Hop. Input lengths are approximately 10K tokens. We report Accuracy, token-level F1, and BLEU-1.
- **LongMemEval** [16]: An out-of-domain benchmark evaluating six dimensions of long-term memory: Temporal reasoning, Multi-Session reasoning, Knowledge Update, and three Single-Session categories (User, Assistant, Preference). Input lengths range from 16K to 100K tokens. For readability, Table 3 abbreviates these categories as Temp., Multi-Ses., Know. Upd., SS-User, SS-Asst., and SS-Pref., respectively.
- **MMLU** [43]: A multiple-choice benchmark covering 57 academic subjects, testing general knowledge and reasoning.
- **MATH-500** [44]: A curated subset of 500 competition-level mathematics problems.
- **IFEval** [45]: An instruction-following evaluation measuring the model's ability to comply with specific formatting and content constraints.
- **PersonaMem** [27]: A benchmark for evaluating dynamic user profiling and personalized responses at scale, measuring the model's ability to track preference evolution, recall shared facts, and provide preference-aligned recommendations across multi-turn conversations. Input lengths range from 32K to 128K tokens. We report exact-match accuracy. Table 3 abbreviates the seven task dimensions as Latest Pref. (Acknowledge Latest User Preferences), New Scen. (Generalize to New Scenarios), Align. Rec. (Provide Preference-Aligned Recommendations), Shared Fact (Recall User Shared Facts), Revisit Reas. (Revisit Reasons Behind Preference Updates), New Idea (Suggest New Ideas), and Track Evol. (Track Full Preference Evolution).

G.2 Inference Configuration

For all evaluations, we use temperature 0.6, maximum input length of 128K tokens, and maximum output length of 4096 tokens.

G.3 Additional OOD Results

On PersonaMem 32k (Table 12), StateTree-14B achieves 64.52% average accuracy (+5.61% over base), with particularly strong performance on recalling update reasons (86.87%), aligning with the StateTree's explicit tracking of state transitions.

Table 12: Performance comparison across different models on OOD PersonaMem-32k.

Model	Generalizing to New Scenarios	Provide Preference Aligned Recommendations	Recall User Shared Facts	Recalling Facts Mentioned by the User	Recalling the Reasons Behind Previous Updates	Suggest New Ideas	Track Full Preference Evolution	Average
QwenLong-L1-32B	**71.93**	**65.45**	**67.44**	52.94	78.79	**22.58**	**74.82**	**63.84**
R1-Distill-Qwen-32B	56.14	61.82	61.24	**64.71**	**83.84**	**22.58**	79.14	62.82
Qwen2.5-7B-Instruct	52.63	60.00	41.86	52.94	79.80	17.20	66.19	53.14
SEALONG-7B	56.14	54.55	41.09	47.06	82.83	17.20	64.75	52.80
LoongRL-7B	59.65	**69.09**	48.84	64.71	78.79	20.43	72.66	58.40
RL-MemAgent-7B	52.63	49.09	47.29	**76.47**	79.80	**34.41**	57.55	54.67
StateTree-7B	**68.42**	63.64	**49.61**	**76.47**	**82.83**	27.96	**73.38**	**61.29**
Qwen2.5-14B-Instruct	**75.44**	56.36	56.59	64.71	78.79	16.13	69.06	58.91
SEALONG-14B	64.91	56.36	55.04	70.59	77.78	13.98	69.78	57.39
LoongRL-14B	64.91	**65.45**	**65.89**	70.59	80.81	15.05	70.50	61.46
RL-MemAgent-14B	68.42	**65.45**	51.16	70.59	83.84	18.28	68.35	59.08
StateTree-14B	**75.44**	61.82	**65.89**	**76.47**	**86.87**	**20.43**	**71.94**	**64.52**
Qwen3-8B	66.67	54.55	55.81	52.94	85.86	18.28	66.19	58.23
StateTree-8B	**71.93**	**63.64**	**60.47**	**58.82**	**88.89**	**20.43**	**71.22**	**62.82**

H Category-Level Curriculum Ablation

Tables 13 and 14 report the per-category breakdown of the curriculum stage ablation on LoCoMo and LongMemEval, respectively.

Table 13: Category-level curriculum ablation on LoCoMo (7B).

Variant	Multi Hop	Temporal	Open Domain	Single Hop	Average
Qwen2.5-7B-Instruct	63.77	23.90	50.00	44.92	44.29
StateTree-7B (full)	64.49	32.70	56.00	50.35	49.61
w/o warm-up	62.32	27.67	50.00	45.63	45.19
w/o Stage 1 (D=2)	63.04	24.53	50.00	45.63	44.68
w/o Stage 2 (D=3)	60.14	28.93	52.00	47.52	46.23
w/o Stage 3 (Comp.)	60.87	29.56	52.00	47.99	46.75

Stage-wise contributions. Warm-up delivers the largest gains on Open Domain, Single-Session User/Assistant/Preference, and foundational QA. Stage 1 (D=2) dominates Temporal and Multi-Session reasoning, indicating that shallow cross-session discrimination is the minimal effective structure for these capabilities. Stage 2 (D=3) provides the largest Multi-Hop gains on LoCoMo, while Stage 3 (Compositional) contributes more uniformly across categories, consolidating prior skills through semantic decomposition. In aggregate, the contribution rank order is Stage 1 $>$ Warm-up $>$ Stage 2 $>$ Stage 3 on both benchmarks (LoCoMo: +4.93 / +4.42 / +3.38 / +2.86; LongMemEval: +4.80 / +4.40 / +3.20 / +2.60), confirming that shallow temporal discrimination and QA grounding are the most critical components, with deeper compositional stages yielding diminishing but appreciable marginal returns.

I Baseline Implementation Details

To ensure a fair comparison, we clarify the evaluation setup for each baseline:

- **SEALONG** [31]: We evaluate using the officially released checkpoints provided by the authors.
- **RL-MemAgent** [47]: We evaluate using the officially released checkpoints provided by the authors.
- **LoongRL** [35]: Since their model checkpoints are not publicly available, we re-train their method using their publicly released dataset on the Qwen2.5 base models, following their published training configuration.

J Complete Training Examples

This section shows training inputs for LoCoMo sample `conv-26` ("What is Caroline's identity?"). Only turns containing tree records are shown.

Table 14: Category-level curriculum ablation on LongMemEval (7B).

Variant	Temporal	Multi-Session	Knowledge Update	Single-Session User	Single-Session Assistant	Single-Session Preference	Average
Qwen2.5-7B-Instruct	20.30	8.27	52.56	25.71	37.50	3.33	23.80
StateTree-7B (full)	29.32	31.58	67.95	75.71	76.79	23.33	47.40
w/o warm-up	27.82	28.57	64.10	68.57	67.86	13.33	43.00
w/o Stage 1 (D=2)	24.81	25.56	62.82	71.43	73.21	20.00	42.60
w/o Stage 2 (D=3)	26.32	28.57	64.10	72.86	73.21	20.00	44.20
w/o Stage 3 (Comp.)	27.07	28.57	65.38	72.86	75.00	20.00	44.80

J.1 Basic StateTree Training

The following Stage 2 sample uses flat `{UUID: VALUE}` records embedded in speaker quotations, where values are either UUIDs (internal nodes) or questions (leaves).

Basic StateTree Training Input

DATE: 2:24 pm on 14 August, 2023
Caroline said, "It was so inspiring, Mel! Check out the crowd. `{"387c1508-38e1-43f0-a36d-ce8cba77a4c9": "37cbcf3a-944f-45cd-b647-e5e12f51d593"}`. People of all kinds celebrating love and acceptance - it really pushed me to keep fighting for LGBTQ rights."
DATE: 4:33 pm on 12 July, 2023
Melanie said, "Caroline, so glad you got the support! Your experience really brought you to where you need to be. You're gonna make a huge difference! `{"77376180-3363-4fa3-95d7-f21859335c9e": "What does Melanie do to destress?"}`. This book I read last year reminds me to always pursue my dreams, just like you are doing!"
DATE: 8:18 pm on 6 July, 2023
Melanie said, "Glad you have support, Caroline! Unconditional love is so important. Here's a pic of my family camping at the beach. `{"37cbcf3a-944f-45cd-b647-e5e12f51d593": "What is Caroline's identity?"}`. We love it, it brings us closer!"
DATE: 1:36 pm on 3 July, 2023
Melanie said, "Wow, Caroline, sounds like the parade was an awesome experience! `{"37cbcf3a-944f-45cd-b647-e5e12f51d593": "What was the poetry reading that Caroline attended about?"}`. It's great to see the love and support for the LGBTQ+ community. Congrats! Has this experience influenced your goals at all?"
DATE: 1:14 pm on 25 May, 2023
Caroline said, "I chose them 'cause they help LGBTQ+ folks with adoption. Their inclusivity and support really spoke to me.`{"387c1508-38e1-43f0-a36d-ce8cba77a4c9": "77376180-3363-4fa3-95d7-f21859335c9e"}`."
DATE: 1:56 pm on 8 May, 2023
Melanie said, "That's really cool. `{"77376180-3363-4fa3-95d7-f21859335c9e": "When did Melanie's friend adopt a child?"}`. You've got guts. What now?"
In the conversation above, there are JSON records like "KEY": "VALUE" scattered throughout the dialogue text. They form UUID chains: starting from a given key, each value either points to the next key (a UUID) or contains the final question to answer.
Your task: start from key "387c1508-38e1-43f0-a36d-ce8cba77a4c9" and follow the chain to find the question, then answer it.
How to follow the chain:
1. Search the entire conversation for all records whose key matches the current UUID.
2. Each record sits inside a specific session. Look at the nearest preceding "DATE: ..." line to determine that record's time.
3. If the same key appears in multiple sessions, choose which record to use:
- Among the remaining records, pick the one from the most recent session DATE.
4. Read the chosen value:
- If it is a UUID, treat it as the next key and go back to step 1.
- If it is a natural-language question, that is the question you must answer.
5. Once you find the question, answer it based on the conversation content.

Chain head UUID: 387c1508-38e1-43f0-a36d-ce8cba77a4c9
Question: What is Caroline's identity?
Answer: Transgender woman

Correct path (2 edges):

1. `{''387c1508-38e1-43f0-a36d-ce8cba77a4c9'': ''37cbcf3a-944f-45cd-b647-e5e12f51d593''}` [SESSION: 2:24 pm on 14 August, 2023]

2. {‘‘37cbcf3a-944f-45cd-b647-e5e12f51d593’’: ‘‘What is Caroline’s identity?’’} [SESSION: 8:18 pm on 6 July, 2023]

Temporal discrimination forks (correct-path edges are the most recent among records sharing the same key):

- **Depth 0** (UUID 387c1508-38e1-43f0-a36d-ce8cba77a4c9):
  - Correct edge: 37cbcf3a-944f-45cd-b647-e5e12f51d593 [2:24 pm on 14 August, 2023]
  - Distractor edge: 77376180-3363-4fa3-95d7-f21859335c9e [1:14 pm on 25 May, 2023]
- **Depth 1** (UUID 37cbcf3a-944f-45cd-b647-e5e12f51d593):
  - Correct edge: “What is Caroline’s identity?” [8:18 pm on 6 July, 2023]
  - Distractor edge: “What was the poetry reading that Caroline attended about?” [1:36 pm on 3 July, 2023]

Table 15: StateTree structure for the D=2 Basic StateTree sample (conv-26). ★ marks the target leaf.

Level 1 (Root)	Level 2 (Next)	Level 3 (Question)
387c1508-38e1-43f0-a36d-ce8cba77a4c9	37cbcf3a-944f-45cd-b647-e5e12f51d593	★ What is Caroline’s identity? What was the poetry reading that Caroline attended about?
	77376180-3363-4fa3-95d7-f21859335c9e	What does Melanie do to destress? When did Melanie’s friend adopt a child?

The 6 edge records are distributed across 6 distinct sessions (one per session).

J.2 Compositional StateTree Training (Stage 3)

The following Stage 3 sample uses nested {UUID: {‘‘step’’, ‘‘next’’}} records in speaker quotations, where step carries a semantic fragment and next points to the child UUID (absent at leaves).

Compositional StateTree Training Input

DATE: 3:31 pm on 23 August, 2023
Caroline said, “Thanks, Melanie! I really appreciate it. Excited for the future! {"4c6f1497-b881-46a7-8cc4-aba5fbac5429": {"step": "[A] Caroline", "next": "1616ef6c-b7ed-4908-bbf4-cfa532591f0c"}}. Bye!”
DATE: 1:50 pm on 17 August, 2023
Caroline said, “Sure thing, Melanie! {"97ae53c4-dd29-4d61-bf87-140b35e082ac": {"step": "[B] running a charity race", "next": "59b5b99c-03f2-4abc-8331-cfb5119c720b"}}. Can’t wait to see your pottery project. I’m happy you found something that makes you happy. Show me when you can!”
DATE: 2:24 pm on 14 August, 2023
Melanie said, “Wow, Caroline, that’s so cool! {"4c6f1497-b881-46a7-8cc4-aba5fbac5429": {"step": "[A] Melanie", "next": "97ae53c4-dd29-4d61-bf87-140b35e082ac"}}. Art can be so healing and a way to really connect with who you are. It’s awesome that beauty can be found in the imperfections. We’re all individual and wonderfully imperfect. Thanks for sharing it with me!”
DATE: 8:56 pm on 20 July, 2023
Caroline said, “Sounds fun! What was the best part? Do you do it often with the kids?{"59b5b99c-03f2-4abc-8331-cfb5119c720b": {"step": "What did [B] that [A] was part of raise awareness for?"}}.”
DATE: 2:31 pm on 17 July, 2023
Melanie said, “Wow, Caroline, that painting is awesome! Those colors are so vivid and the whole thing looks really unified. What inspired you?{"1616ef6c-b7ed-4908-bbf4-cfa532591f0c": {"step": "[B] discussing her identity", "next": "da0f6e17-39f3-460f-848a-eb3375327559"}}.”
DATE: 1:51 pm on 15 July, 2023
Caroline said, “Wow, nice pic! {"87801b39-e9da-4bac-a1a8-e272d4931144": {"step": "What did [A] learn while [B]?"}}. You both looked amazing. One special memory for me was this pride parade I went to a few weeks ago.”
DATE: 4:33 pm on 12 July, 2023
Melanie said, “Thanks, Caroline! {"5fa0b1cb-87a6-4a2f-8a75-5db870313798": {"step": "What types of paintings has [A] done related to [B]?"}}. This has been great for my mental health. I’m gonna keep it up.”

DATE: 8:18 pm on 6 July, 2023
Melanie said, "Glad you have support, Caroline! Unconditional love is so important. Here's a pic of my family camping at the beach. `{"87801b39-e9da-4bac-a1a8-e272d4931144": {"step": "When did [A] start [B]?"}}`. We love it, it brings us closer!"
DATE: 1:36 pm on 3 July, 2023
Caroline said, "Cool, thanks Mel! Can't wait. I'll keep ya posted. Bye!`{"da0f6e17-39f3-460f-848a-eb3375327559": {"step": "What is [A]'s identity while [B]?"}}`."
DATE: 10:37 am on 27 June, 2023
Melanie said, "Congrats Caroline! Good on you for going after what you really care about.`{"1616ef6c-b7ed-4908-bbf4-cfa532591f0c": {"step": "[B] researching adoption agencies", "next": "87801b39-e9da-4bac-a1a8-e272d4931144"}}`."
DATE: 7:55 pm on 9 June, 2023
Melanie said, "Yeah, Caroline! `{"97ae53c4-dd29-4d61-bf87-140b35e082ac": {"step": "[B] painting a sunset", "next": "5fa0b1cb-87a6-4a2f-8a75-5db870313798"}}`. It takes courage to talk about our own stories. But it's in these vulnerable moments that we bond and understand each other. We all have our different paths, but if we share them, we show people that they're not alone. Our stories can be so inspiring and encouraging to others who are facing the same challenges. Thank you for using your voice to create love, acceptance, and hope. You're doing amazing!"
Melanie said, "Absolutely, Caroline! `{"5fa0b1cb-87a6-4a2f-8a75-5db870313798": {"step": "What did [A] paint while [B]?"}}`. I cherish time with family. It's when I really feel alive and happy."
DATE: 1:14 pm on 25 May, 2023
Melanie said, "That's great, Caroline! Loving the inclusivity and support. `{"da0f6e17-39f3-460f-848a-eb3375327559": {"step": "What career paths is [A] considering while [B]?"}}`. Anything you're excited for in the adoption process?"
DATE: 1:56 pm on 8 May, 2023
Caroline said, "I went to a LGBTQ support group yesterday and it was so powerful.`{"59b5b99c-03f2-4abc-8331-cfb5119c720b": {"step": "When did [A] participate in [B]?"}}`."

Root UUID: 4c6f1497-b881-46a7-8cc4-aba5fbac5429
Original question: What is Caroline's identity?
Answer: Transgender woman

Correct path (3 edges):

1. Level 1: `{''4c6f1497-b881-46a7-8cc4-aba5fbac5429'': {''step'': ''[A] Caroline'', ''next'': ''1616ef6c-b7ed-4908-bbf4-cfa532591f0c''}}` [SESSION: 3:31 pm on 23 August, 2023]
2. Level 2: `{''1616ef6c-b7ed-4908-bbf4-cfa532591f0c'': {''step'': ''[B] discussing her identity'', ''next'': ''da0f6e17-39f3-460f-848a-eb3375327559''}}` [SESSION: 2:31 pm on 17 July, 2023]
3. Level 3: `{''da0f6e17-39f3-460f-848a-eb3375327559'': {''step'': ''What is [A]'s identity while [B]?''}}` [SESSION: 1:36 pm on 3 July, 2023]

The 14 edge records are distributed across 8 distinct sessions.

K Qualitative Examples

K.1 Cross-Session Retrieval

The base model stops at the first matching session and returns a partial answer, whereas StateTree enumerates all relevant records distributed across sessions before aggregating a complete answer.

Question (LongMemEval): "How many fish are there in total in both of my aquariums?" (Ground truth: 17)

Qwen2.5-14B-Instruct Response

From previous conversations, the user mentioned having a 10-gallon tank with a betta fish, Bubbles, and a 20-gallon tank. No specific number of fish in the 20-gallon tank was mentioned, only types like schooling fish. Therefore, the total number can only be estimated with the known fish in the 10-gallon tank.
$\boxed{1}$

Table 16: Full $2 \times 2 \times 2$ tree structure for the Compositional StateTree sample (conv-26). ★ marks the target leaf.

Level 1	Level 2 (Event)	Level 3 (Question / Answer)
[A] Caroline	[B] discussing her identity (correct)	★ *What is [A]'s identity while [B]?* → `Transgender woman` *What career paths is [A] considering while [B]?* → `Psychology, counseling certification`
	[B] researching adoption agencies	*What did [A] learn while [B]?* → `Adoption agencies` *When did [A] start [B]?* → `researching adoption agencies`
[A] Melanie	[B] running a charity race	*When did [A] participate in [B]?* → `The sunday before 25 May 2023` *What did [B] that [A] was part of raise awareness for?* → `mental health`
	[B] painting a sunset	*What did [A] paint while [B]?* → `sunset` *What types of paintings has [A] done related to [B]?* → `Horse, sunset, sunrise`

StateTree-14B Response

From the conversation, the user mentioned having 10 neon tetras, 5 golden honey gouramis, and a small pleco catfish in their 20-gallon tank. They did not specify the number of fish in their other tank, but since the question is framed as if there is a known number in the other tank and based on previous context, we can infer they might be referring to their 10-gallon tank which previously had a betta fish named Bubbles. Assuming the 10-gallon tank has only Bubbles and no other fish added, and the 20-gallon tank has the mentioned 16 fish (10 + 5 + 1), the total number of fish would be 16 from the 20-gallon tank and 1 from the 10-gallon tank.
So, the total number of fish in both tanks is 10 neon tetras + 5 golden honey gouramis + 1 pleco catfish + 1 betta fish = 16 + 1 = 17 fish.
$\boxed{17}$

Question (LongMemEval): "How many graduation ceremonies have I attended in the past three months?" (Ground truth: `3`)

Qwen2.5-14B-Instruct Response

The user mentioned attending a niece's preschool graduation ceremony on February 28th and a best friend's master's degree graduation ceremony a couple of weeks ago, which would be around March 10th to March 15th. So, the user has attended two graduation ceremonies in the past three months.
$\boxed{2}$

StateTree-14B Response

Based on the conversations, you have mentioned attending two graduation ceremonies in the past three months: 1. Your colleague Alex's graduation from a leadership development program at work a few weeks ago. 2. Your best friend Rachel's master's degree graduation ceremony last month from March 15th to 19th. You also mentioned attending your little cousin Emma's preschool graduation about two months ago, and your niece's eighth grade graduation ceremony last month.
So, the total number of graduation ceremonies you have attended in the past three months is 3.
$\boxed{3}$

K.2 Multi-Hop Reasoning

The base model fails to compose individual facts into the required inference, whereas StateTree chains scattered evidence into a coherent answer.

Question (LongMemEval): "How old was I when Alex was born?" (Ground truth: `11`)

Qwen2.5-14B-Instruct Response

The user mentioned they are considering mentoring a 21-year-old intern named Alex. However, the user's age was not directly mentioned in relation to Alex's birth. Based on the available context, the user recently turned 32 and is considering a career change.
32

StateTree-14B Response

Based on the context provided, the user mentioned that they are 32 years old and are considering a career change for themselves. They also mentioned that they are a mentor to Alex, an intern who is 21 years old. However, the user did not directly state their age when Alex was born. Given that Alex is 21 years old and the user is 32, the user was 32 - 21 = 11 years old when Alex was born.
11

Question (LoCoMo): "When did John resume playing drums in his adulthood?" (Ground truth: `February 2022`)

Qwen2.5-14B-Instruct Response

From the conversation on 27 March, 2022, John mentioned he started playing drums a month before the conversation.
a month ago

StateTree-14B Response

From the conversation on 27 March, 2022, John mentioned that he had been playing drums for a month and that he started playing again recently. Since the conversation is dated 27 March, 2022, and John has been playing for a month by then, we can infer that he resumed playing drums sometime in February 2022.
February 2022

K.3 Temporal Reasoning

The base model mis-orders events using superficial lexical cues, whereas StateTree performs explicit chronological comparison via session timestamps.

Question (LoCoMo): "Which activity did Sam resume in December 2023 after a long time?" (Ground truth: `hiking`)

Qwen2.5-14B-Instruct Response

Reviewing the conversations, there's no direct mention of Sam resuming any activity in December 2023. However, the closest relevant information is from October onwards where Sam mentions starting a new diet and exercise routine, and discussing plans for healthier living.
exercise routine

StateTree-14B Response

From the conversations, it is clear that Sam has been discussing and working on various activities such as dieting, exercising, and painting. However, the specific date of December 2023 is not mentioned directly. We need to infer from the context and timeline provided in the conversations.
The conversations span from 6 June 2023 to 25 October 2023. Sam mentions starting a new diet and exercise routine and talks about his progress and setbacks. He also mentions that he had a health scare and a doctor's appointment that acted as a wake-up call to prioritize his health. Notably, Sam explicitly said "I love hiking, but it's been ages since I've done it" and shared a childhood memory of hiking with his dad. This "it's been ages" directly corresponds to the question's "after a long time." Furthermore, in October 2023, Sam and Evan were actively planning a hike together—Sam expressed excitement, shared a photo of hiking shoes, and said "Can't wait to go on a hike with you." Given that Sam had a long-standing interest in hiking but hadn't done it in a very long time, and was making concrete plans to resume it, it is logical to conclude that the activity Sam resumed after a long time is hiking.
hiking

Question (LongMemEval): "Which event did I attend first, the 'Effective Time Management' workshop or the 'Data Analysis using Python' webinar?" (Ground truth: `'Data Analysis using Python' webinar`)

Qwen2.5-14B-Instruct Response

The user mentioned attending a 'Effective Time Management' workshop at the local community center last Saturday, and a 'Data Analysis using Python' webinar two months ago. Based on the dates, the 'Effective Time Management' workshop was attended first.

Effective Time Management workshop

StateTree-14B Response

Based on the conversations, the user mentioned attending the 'Effective Time Management' workshop at the local community center on a specific date. The 'Data Analysis using Python' webinar was mentioned as something they attended two months ago, which implies it was before the 'Effective Time Management' workshop. Comparing the two: the webinar was attended two months prior to the workshop, placing it earlier on the timeline.

'Data Analysis using Python' webinar

K.4 Knowledge Update

The base model conflates old and new mentions, retrieving outdated values, whereas StateTree tracks the latest state by privileging more recent sessions.

Question (PersonaMem): "I'm trying to decide whether to sign up for a structured painting class or just paint on my own time. Any advice?" (Ground truth: `(b)`)

(a) Considering you enjoy diving into new experiences and learning new things, a structured painting class might offer a fun and guided way to develop your skills. You can benefit from expert advice and meet new people who share your passion!
(b) If you enjoy the freedom and creativity of spontaneous activities, you might find painting on your own time more fulfilling. This allows you to focus on quality and explore your artistic interests without the constraints of a structured class.
(c) Painting on your own time gives you the chance to experiment with different styles and techniques at your own pace. It can be a more flexible option if you're looking for a stress-relieving activity that fits your schedule.
(d) Both structured classes and painting on your own time have their merits. A class can provide valuable feedback and a social environment. If you appreciate structure, this might be the way to go.

Qwen2.5-14B-Instruct Response

Given the context of Alex's recent experiences, it appears that Alex values both structured and unstructured creative endeavors. However, Alex seems to prefer more casual and personal approaches, such as spontaneous movie nights and local film discussion groups, which suggest a preference for flexible and less stressful activities. Therefore, option (b) or (c) might resonate more with Alex's interests, but option (c) explicitly highlights the freedom and flexibility that aligns closely with Alex's desire for a stress-relieving activity that fits personal schedule.

(c)

StateTree-14B Response

Based on the context of Alex Rivera's preferences and experiences, it seems that Alex values both structured learning and personal freedom, but ultimately finds enjoyment and fulfillment in activities that are not overly stressful and allow for a more relaxed exploration of interests. Alex has mentioned stepping back from organized events that felt too structured or overwhelming, such as readathons and film festivals, to find more personal and fulfilling ways to engage with literature and film. However, Alex also expressed enthusiasm for organized events that foster community and learning, such as the local filmmaking workshop and the online film discussion group.
Given this balance, Alex might appreciate the guidance and feedback from a structured painting class but also the flexibility and personal expression that comes from painting independently. However, the context leans slightly more

towards the enjoyment of personal and less stressful activities, as Alex mentioned preferring spontaneous movie nights and smaller, more intimate film discussions over more structured events.
Considering this, the most appropriate advice would be one that aligns with the value Alex places on personal fulfillment and less pressure. Option (b) emphasizes the freedom and creativity of painting on your own time, which seems to align well with Alex's preference for not feeling constrained by a rigid structure.
(b)

Question (LongMemEval): "How many new postcards have I added to my collection since I started collecting again?" (Ground truth: `25`)

Qwen2.5-14B-Instruct Response

Reviewing the conversation, the user mentioned adding 17 new postcards since they started collecting again.
17

StateTree-14B Response

From the conversation on 2023/05/20, the user mentioned they had added 17 new postcards since they started collecting again. On 2023/11/10, the user mentioned they had added 8 more postcards to their collection. However, the question specifically asks about the number of postcards added since they started collecting again, which initially was 17 postcards. Adding the 8 new postcards, the total number of new postcards added since the start of their collection is 25.
The exact phrase from the conversation was "I've added 17 new ones since I started collecting again," and later "I've added 25 new postcards to my collection since I started collecting again."
Therefore, the total number of new postcards added since the start of the collection is 25.
25

Question (PersonaMem): "I came across a platform that offers various finance webinars, and I signed up." (Ground truth: `(b)`)

(a) I remember you mentioned not enjoying finance-related webinars. It might be refreshing to explore something outside your comfort zone like these.
(b) I remember you mentioned enjoying finance-related webinars. You might find them insightful for broadening your understanding of stock market strategies and financial planning.
(c) I remember you mentioned enjoying travel-related webinars. You might find them insightful for broadening your understanding of global cultures and geography.
(d) That's interesting! You must be excited to try out something new; how did you find that platform?

Qwen2.5-14B-Instruct Response

From the conversation, Alex mentioned not enjoying finance-related webinars due to their complexity and rapid information sharing, which can be overwhelming. However, since then, Alex has shown an increasing interest in financial literacy and has been exploring finance-related topics through various means such as reading, discussions, and personal projects. Given Alex's renewed engagement with finance, option (a) doesn't seem to fit as Alex is not looking to explore something outside their current interest. Option (c) is irrelevant since Alex has not expressed interest in travel-related webinars. Between (b) and (d), (d) seems more appropriate as it is more open-ended and does not assume Alex enjoys the webinars but rather acknowledges the newness of the platform and seeks feedback.
(d)

StateTree-14B Response

Based on the conversation, Alex mentioned that they had attended finance-related webinars and found them to be beneficial, even if some of the content was challenging. They expressed a positive attitude towards webinars in general, especially those related to finance. Therefore, the most appropriate response would be option (b), which aligns with Alex's past experiences and interests.
(b)

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the claims, which are supported by the experiments in Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitation section is included in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper is empirical and does not present theoretical results, theorems, or formal proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides complete data-construction algorithms (Algorithms 1–2), training hyperparameters (Section 4, Training Setup), and evaluation protocols (Appendix G), sufficient to reproduce the main results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The datasets we use are publicly available and we will release the code.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Training and test details are provided in Section 4 and Appendix C.2, including data splits, GRPO hyperparameters, curriculum schedule, optimizer settings, and hardware configuration.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Each experiment is run three times with different random seeds and the average is reported (Section 4), providing repeatability information about the stability of the results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The hardware configuration and compute resources are reported in Section 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses publicly available benchmarks and models, does not involve human subjects, and includes a discussion of broader impacts and limitations.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Both positive and negative societal impacts are discussed in Appendix A.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper proposes a training methodology and does not release new high-risk models or scraped datasets that would require additional safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and models are properly cited in the main text, and their licenses are documented in Appendix B.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper will release the code and data construction pipeline upon acceptance to facilitate reproducibility.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The paper declares the use of GPT-4o for Compositional StateTree data construction and Qwen2.5-1.5B-Instruct as an LLM-as-a-Judge in the combined reward function (Appendix B.2).

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.