THINK ONTOLOGICALLY: INTERNALIZING NEURO-FORMAL STATE INVARIANTS FOR ROBUST CONVERSATIONAL MEMORY IN LARGE LANGUAGE MODEL AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model (LLM) agents operating in multi-turn dialogues suffer from severe conversational memory degradation: external context compression and prose summarization trigger continuous semantic drift, whereas detached symbolic reasoners impose non-differentiable foreign-function bottlenecks and rigid extraction barriers. In this work, we propose **Think Ontologically**, a neuro-formal reasoning framework that internalizes conversational state tracking directly into neural policy weights. Rather than offloading memory to external solvers or prose buffers, the agent autoregressively synthesizes first-order relational state invariants (`causes`, `holds_value`, `has_state`, `has_goal`) grounded in foundational ontologies (UFO and BFO) before generating conversational responses. We train foundation models using a two-stage curriculum: supervised warmup on verified relational tuples followed by Verifiable Reinforcement Fine-Tuning (RLVR) with deterministic multi-tiered reward signals. Investigating multi-reward credit assignment dynamics, we demonstrate that Group reward-Decoupled Normalization Policy Optimization (GDPO) and sequence-level importance sampling (GSPO) effectively eliminate multi-reward gradient masking, enabling models to satisfy deductive constraints while maintaining 100% syntax validity and minimal policy drift ($\mathbb{D}_{\text{KL}} \leq 0.0027$). Across five open-weight model families (Qwen-2.5-14B/7B, Llama-3.1-8B, Gemma-3-4B, and DeepSeek-R1-Distill-7B), our internalized policy consistently outperforms zero-shot and in-context baselines, boosting IFEval strict instruction compliance by up to $+6.63$ percentage points, elevating GPQA Diamond scientific reasoning to 29.80%, and achieving an 88.0% goal completion rate on τ-bench Airline interactions with a $> 75\%$ reduction in inference latency compared to decoupled neuro-symbolic pipelines.

1 INTRODUCTION

Transformer-based Large Language Model (LLM) agents (Vaswani et al., 2017) deployed in conversational and task-oriented settings must maintain reliable internal state representations across extended dialogue trajectories (Guan et al., 2026). When dialogue histories span multiple conversational turns, models encounter what we term the **Conversational Memory Trilemma**: an architectural tradeoff between (1) **Context Saturation**, where storing complete interaction histories exhausts the finite attention context window, incurring quadratic computational overhead and triggering "lost-in-the-middle" degradation (Liu et al., 2024); (2) **Semantic Drift**, where compressing past interactions via recursive natural language summarization causes progressive entropy and cumulative condition erosion, discarding subtle constraints (e.g., negative instructions, transient states, and numerical thresholds) (Carranza & Rojas, 2025; Verma, 2026); and (3) **Decoupled Solver Barriers**, where delegating memory management to detached symbolic reasoning engines (e.g., Prolog, Datalog, or SMT solvers) introduces rigid two-stage extraction bottlenecks, non-differentiable execution boundaries, syntax fragility, and multi-second foreign-function interface (FFI) latencies (Pan et al., 2023; Olausson et al., 2023).

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

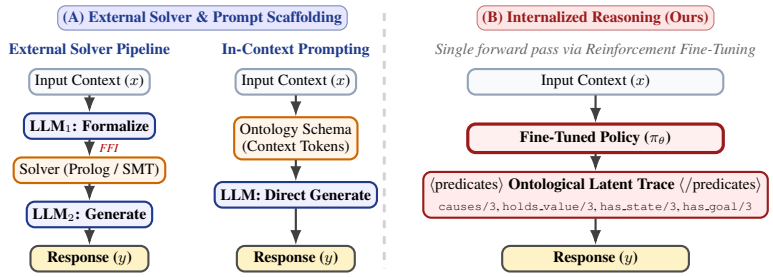

Figure 1: Comparison of conversational state tracking and reasoning architectures. **(A) External solver and prompt-scaffolding pipelines**: Suffer from multi-stage extraction bottlenecks, syntax fragility, and non-differentiable foreign-function barriers (FFIs). **(B) Internalized neuro-formal reasoning (Ours)**: End-to-end policy optimization where relational ontological state invariants are synthesized directly in an autoregressive scratchpad, achieving robust state tracking in a single forward pass without runtime dependencies.

Existing approaches to conversational memory fall into two mutually flawed extremes: either treating memory as malleable, ungrounded prose (Packer et al., 2023; Chhikara et al., 2025) or offloading state tracking to detached symbolic engines (Pan et al., 2023; Gao et al., 2023). Historically, neuro-symbolic research assumed that formal soundness requires external symbolic solvers, imposing non-differentiable FFI barriers and execution latency. Conversely, conversational agent research assumed memory must be stored as soft prose buffers, inevitably succumbing to semantic drift and lost-in-the-middle degradation over long horizons.

To resolve this trilemma, we introduce **Think Ontologically**. We posit that dialogue state should be modeled neither as malleable prose nor as a detached external database, but as *typed formal ontological invariants* synthesized directly within the neural policy's autoregressive reasoning stream. In extended multi-turn interactions, conversational context naturally decomposes into four foundational dimensions: (1) enduring entity properties and user preferences that must persist across turns, (2) transient situational states that capture the immediate interaction phase, (3) teleological goals that define the user's intended outcomes, and (4) causal events that govern valid state transitions. Grounding these dimensions in Unified Foundational Ontology (UFO) and Basic Formal Ontology (BFO) (Guizzardi, 2005; Guizzardi et al., 2022; Arp et al., 2015), our model learns to emit structured relational state assertions in a compact scratchpad before generating each response. Using Verifiable Reinforcement Fine-Tuning (RLVR) with Group Relative Policy Optimization (GRPO) (Shao et al., 2024), the policy internalizes formal state tracking directly into its forward pass, eliminating external solver overhead while guaranteeing deductive consistency (Figure 1).

Contributions. Resolving each horn of the Conversational Memory Trilemma, our contributions are:

- **Resolving Semantic Drift via Invariants**: We formalize dialogue state tracking as typed first-order relational invariants grounded in foundational ontologies (UFO/BFO) in the scratchpad, proving reward optimality guarantees constraint satisfaction (Proposition 1).
- **Resolving Solver Barriers via RLVR**: We formulate a deterministic, multi-tiered reward suite verifying delimiter syntax, schema conformance, causal DAG acyclicity, and task constraints without external provers or LLM judges.
- **Overcoming Gradient Masking via GDPO**: We analyze credit assignment across GRPO, GSPO (Zheng et al., 2025), and GDPO (Liu et al., 2026), showing that decoupled advantage normalization prevents formatting rewards from masking deductive task accuracy.
- **Long-Context Empirical Gains**: Across 5 foundation model families on IFEval, GPQA Diamond, and τ-bench, our internalized policy prevents lost-in-the-middle degradation (84.0% constraint preservation, 88.0% goal completion) with $> 75\%$ lower latency than external pipelines.

2 Related Work

2.1 Conversational Memory and Context Compression in LLMs

Early conversational agents utilized sliding context windows or recursive summarization buffers to handle long dialogue histories (Packer et al., 2023; Zhong et al., 2024). Retrieval-augmented generation (RAG) architectures (Lewis et al., 2020) retrieve relevant passages from external stores, but vector-similarity retrieval surfaces disjointed fragments failing to capture temporal state updates or causal transitions (Jiang et al., 2023; Verma, 2026). Systems such as Mem0 (Chhikara et al., 2025), A-MEM (Xu et al., 2025), and MemMA (Lin et al., 2026) employ hierarchical memory banks, while Reflexion (Shinn et al., 2023) and Self-Refine (Madaan et al., 2023) use verbal feedback for iterative correction. Recent efforts explore self-supervised context memory training (MemTrain) (Li et al., 2026b) and lightweight cognitive memory management (LiCoMemory) (Huang et al., 2025). Yet, representing memory as unstructured prose leaves them susceptible to semantic drift over long horizons. In contrast, our approach encodes dialogue history as typed first-order relational invariants with immutable semantics.

2.2 Neuro-Symbolic Reasoning and Formal Solvers

Bridging neural networks with formal logic has long been pursued to guarantee deduction faithfulness across complex reasoning tasks (Olausson et al., 2023). Frameworks such as Logic-LM (Pan et al., 2023) and Program-Aided Language models (PAL) (Gao et al., 2023) prompt LLMs to translate natural language problems into symbolic programs executed by external interpreters (e.g., Z3 (de Moura & Bjørner, 2008), Prolog, Python). SatLM (Ye et al., 2023) converts reasoning tasks to SAT problems, while Faithful CoT (Lyu et al., 2023) generates symbolic programs from chain-of-thought traces. In the domain of formal theorem proving, Lean-STaR (Lin et al., 2025) interleaves informal natural-language thoughts with formal Lean 4 tactics, bootstrapping via expert iteration with compiler verification as reward. While guaranteeing deductive soundness, these decoupled pipelines introduce a rigid two-stage extraction bottleneck: any syntactic or semantic translation error causes catastrophic execution failures. Furthermore, calling external provers across multi-turn interactions introduces significant latency (often 2–5 seconds per turn) and non-differentiable boundaries that prevent end-to-end policy optimization. Our framework internalizes the formal reasoning engine directly into the neural policy weights through rule-based RLVR, achieving deductive consistency in a single, differentiable forward pass without external dependencies.

2.3 Dialogue State Tracking and Formal Ontologies

Dialogue State Tracking (DST) tracks user commitments and goals across multi-turn exchanges (Guan et al., 2026; Carranza & Rojas, 2025). Traditional neural DST architectures map utterances into rigid slot-value tables defined over fixed domain schemas such as MultiWOZ (Budzianowski et al., 2018) and SGD (Rastogi et al., 2020), which fail when conversations shift outside predefined ontologies. In-context learning approaches (Budzianowski et al., 2022) partially address schema rigidity but remain bounded by the prompt's context window. Concurrently, formal ontology engineering (e.g., UFO (Guizzardi et al., 2022) and BFO (Arp et al., 2015)) provides mathematically rigorous metamodels distinguishing enduring entity qualities (`holds_value`) from transient situational moments (`has_state`) and teleological commitments (`has_goal`). Rather than maintaining flat slot tables or querying external provers, *Think Ontologically* internalizes foundational ontological primitives directly into autoregressive policy weights via RLVR, achieving sound state tracking without schema brittleness.

2.4 Reinforcement Learning and Self-Improvement for LLM Reasoning

Tracing back to the scratchpad paradigm (Nye et al., 2021), chain-of-thought (CoT) prompting in few-shot (Wei et al., 2022) and zero-shot (Kojima et al., 2022) regimes, alongside extensions such as self-consistency decoding (Yao et al., 2023b), tree-structured search (Yao et al., 2023a), and interleaved reasoning-action loops (ReAct) (Yao et al., 2023c), establishes that intermediate reasoning tokens improve downstream accuracy but provide no guarantee of formal deductive correctness. Self-improvement methods bootstrap reasoning data from model-generated solutions:

STaR (Zelikman et al., 2022) filters correct rationales for SFT, while Quiet-STaR (Zelikman et al., 2024) trains models to generate implicit thoughts at every token position. Reinforced self-training (ReST/ReST^{EM}) (Gulcehre et al., 2023) extends this loop with reward-weighted sampling. On the policy optimization front, RLHF with PPO (Schulman et al., 2017; Ouyang et al., 2022) remains foundational, while offline alternatives such as DPO (Rafailov et al., 2023) bypass reward model training. In verifiable domains, early work on verifiers (Cobbe et al., 2021), verifier-guided self-training (V-STaR) (Hosseini et al., 2024), automated step-level process reward models (Lightman et al., 2024; Wang et al., 2024a), and group relative policy optimization (GRPO) (Shao et al., 2024; DeepSeek-AI et al., 2025) enable RL with verifiable supervision. Our work extends this line by defining ontological predicate validity as a verifiable reward signal, coupling multi-objective RLVR (GDPO) with foundational ontology to internalize formal state tracking.

3 Methodology: The Think Ontologically Framework

Framework Overview. As illustrated in Figure 1B, *Think Ontologically* internalizes conversational state tracking directly into policy weights via verifiable reinforcement learning. Given dialogue history x, the policy π_θ first autoregressively emits a compact ontological scratchpad z before producing response y. This scratchpad encapsulates dialogue state $\mathcal{S}_t$ as typed first-order relational invariants. During training, policy exploration is steered by an automated rule-based reward suite evaluating delimiter syntax, schema validity, causal graph acyclicity (Directed Acyclic Graph, DAG), and task constraints. Using Group reward-Decoupled Normalization Policy Optimization (GDPO), the model learns to synthesize valid formal state representations in a single forward pass, completely eliminating external logic provers at inference.

3.1 Ontological State Space Formalization

To model conversational context without semantic degradation, we formalize dialogue dynamics using four typed first-order relational predicates rooted in Unified Foundational Ontology (UFO) and Basic Formal Ontology (BFO) (Guizzardi, 2005; Guizzardi et al., 2022; Arp et al., 2015).

Definition 1 (Ontological State Space)**.** *Let $\mathcal{C}$ denote conversational contexts, $\mathcal{E}$ entities, $\mathcal{V}$ constraints, and $\mathcal{G}$ objectives. The conversational state at turn t, denoted $\mathcal{S}_t = \langle \mathcal{R}_{cause}, \mathcal{R}_{value}, \mathcal{R}_{state}, \mathcal{R}_{goal} \rangle$, is a tuple of typed relational assertions: (1) $\mathcal{R}_{cause} \subseteq \mathcal{C} \times \mathcal{E} \times \mathcal{E}$ encodes causal transitions via* `causes`(c, e, e')*; (2) $\mathcal{R}_{value} \subseteq \mathcal{C} \times \mathcal{E} \times \mathcal{V}$ enforces persistent invariants and user constraints via* `holds_value`(c, e, v)*; (3) $\mathcal{R}_{state} \subseteq \mathcal{C} \times \mathcal{E} \times \mathcal{V}$ captures transient situational attributes via* `has_state`(c, e, s)*; and (4) $\mathcal{R}_{goal} \subseteq \mathcal{C} \times \mathcal{E} \times \mathcal{G}$ tracks teleological commitments via* `has_goal`(c, e, g).

These four predicates instantiate core ontological categories from UFO and BFO: `causes/3` models UFO's *Event–Event causation* relation; `holds_value/3` encodes BFO's *specifically dependent continuant* qualities (persistent constraints); `has_state/3` captures UFO's *intrinsic moment* (transient situational attributes); and `has_goal/3` maps to BFO's *objective specification* (intentional teleological commitments).

Why First-Order Relational Predicates? We formulate state assertions strictly as ground First-Order Logic (FOL) predicates over concrete dialogue individuals rather than Second-Order Logic (SOL). Ground FOL model checking on finite domains is decidable in linear time ($O(|z|)$), enabling instantaneous reward verification during RLVR, whereas SOL introduces undecidability and intractable proof search (Appendix F). The closed vocabulary $\mathcal{F}_{\text{onto}}$ precludes hallucinated predicate drift.

During inference, given dialogue history x, policy π_θ autoregressively emits the ontological scratchpad $z = \mathcal{S}_t$ before response y:

```
<start_predicates>
has_goal(ctx, user, "reschedule"). holds_value(ctx, booking, "aisle_seat"). causes(ctx, "
    reschedule", "update_state").
</end_predicates>
<RESPONSE> Your flight has been rescheduled with your preferred aisle seat confirmed. </
    RESPONSE>
```

3.2 Verifiable Rule-Based Reward Formulation (RLVR)

To provide dense, verifiable feedback without human annotators or external solver calls at test time, we construct a deterministic, rule-based reward function $R(q, o) = R_{\text{format}}(o) + R_{\text{schema}}(z) + R_{\text{graph}}(z) + R_{\text{acc}}(q, y)$ over candidate completions $o = (z, y)$:

$$R_{\text{fmt}}(o) = 1.0 \cdot \mathbb{I}(o \in \mathcal{L}_{\text{XML}}) - 1.5 \cdot \mathbb{I}(o \notin \mathcal{L}_{\text{XML}}), \quad R_{\text{graph}}(z) = 1.0 \cdot \mathbb{I}(G_{\text{causes}}(z) \in \text{DAG}), \tag{1}$$

$$R_{\text{schema}}(z) = \sum_{p \in z} 0.5 \cdot \mathbb{I}(\text{fn}(p) \in \mathcal{F}_{\text{onto}} \wedge \text{ar}(p) = 3), \quad R_{\text{acc}}(q, y) = 2.0 \cdot \mathbb{I}(\text{eval}) - 1.0 \cdot \mathbb{I}(\neg\text{eval}), \tag{2}$$

where $\mathcal{L}_{\text{XML}}$ enforces valid XML tags, $\mathcal{F}_{\text{onto}} = \{\texttt{causes}, \texttt{holds_value}, \texttt{has_state}, \texttt{has_goal}\}$, $G_{\text{causes}}(z) \in$ DAG ensures the causal transition graph is a Directed Acyclic Graph (DAG) preventing circular causal contradictions, and $\text{eval}(q, y)$ executes deterministic constraint verification.

3.3 Two-Stage Training Pipeline: Supervised Warmup and RLVR

Training a foundation model to internalize structured ontological invariants presents a cold-start challenge: randomly initialized exploration under RLVR yields near-zero valid scratchpads, resulting in uninformative reward signals. To resolve this, our framework employs a two-stage curriculum:

1. **Stage 1: Supervised Warmup (SFT)**: We bootstrap on $N = 1,500$ verified episodes (2 epochs, lr 2×10^{-5}), priming attention weights with delimiter syntax and predicate arities.
2. **Stage 2: Verifiable RLVR**: Initialized from SFT, π_θ is optimized via group relative policy optimization ($G = 8$, temp $\tau = 0.9$, lr 5×10^{-6}, KL penalty $\beta = 0.04$) to maximize task success.

Both stages use LoRA adapters (Hu et al., 2022) ($r = 64$, $\alpha = 128$, targeting projection matrices; $\sim$0.5% parameters). The $N = 1,500$ seed episodes are drawn from LoCoMo-Plus (Li et al., 2026a; Maharana et al., 2024) and filtered by the verification suite ($R_{\text{schema}} > 0 \wedge R_{\text{graph}} > 0$). All test benchmarks (IFEval, GPQA Diamond, τ-bench) are strictly held-out.

3.4 Policy Optimization Dynamics: GRPO, GSPO, and GDPO

In multi-reward RLVR, standard GRPO (Shao et al., 2024) computes advantages by summing scalar rewards across all objectives:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\left(r_{i,t}(\theta)\hat{A}_i, \text{clip}(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_i\right) - \beta\mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}), \tag{3}$$

where $r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|q, o_{i,<t})}{\pi_{\text{old}}(o_{i,t}|q, o_{i,<t})}$ and $\hat{A}_i = \frac{R_i - \text{mean}(\{R_j\})}{\text{std}(\{R_j\}) + \delta}$. However, standard GRPO exhibits two fundamental failure modes in neuro-formal reasoning:

1. **Token-Level Ratio Volatility**: Autoregressive generation of symbolic delimiters produces high variance. **GSPO** (Zheng et al., 2025) replaces token ratios with geometric sequence-level importance sampling: $s_i(\theta) = \exp\left(\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\log\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\text{old}}(o_{i,t}|q,o_{i,<t})}\right)$.
2. **Multi-Reward Gradient Masking**: Dense formatting rewards dominate variance, rewarding syntax over accuracy. **GDPO** (Liu et al., 2026) normalizes advantages independently per reward dimension: $\hat{A}_i^{(k)} = \frac{R_i^{(k)} - \text{mean}(\{R_j^{(k)}\})}{\text{std}(\{R_j^{(k)}\}) + \delta}$, $\hat{A}_i = \sum_k w_k \hat{A}_i^{(k)}$, maintaining gradient sensitivity to deductive accuracy.

Decoupled normalization guarantees that each reward dimension retains non-zero gradient variance within its group, preventing the easy-to-satisfy formatting reward from dominating the harder task-level deductive accuracy objective.

3.5 Reward-Theoretic Guarantee

Proposition 1 (State Invariant Consistency by Reward Design)**.** *Let* $R(q, o) = R_{format}(o) + R_{schema}(z) + R_{graph}(z) + R_{acc}(q, y)$ *be the composite reward over completions* $o = (z, y)$*. If policy* π_θ *achieves optimal reward* $R^*(q) = \max_o R(q, o)$*, then* y *satisfies all verifiable constraints* $\mathcal{V}_q$ *implied by state* $\mathcal{S}_t$*.*

Proof Sketch. Because $R_{\text{acc}} \in \{-1.0, +2.0\}$ while syntax rewards are bounded ($R_{\text{format}} \leq 1.0$, $R_{\text{graph}} \leq 1.0$), any completion with $R_{\text{acc}} = -1.0$ is strictly dominated by $+3.0$ by a constraint-satisfying candidate. Hence R^* strictly implies $\text{eval}(q, y) = \text{true}$ (see Appendix A for the complete formal proof). □

4 Experimental Evaluation

Foundation Model Suite. We evaluate across five open-weight model families spanning 4B to 14B parameters: Qwen-2.5-14B-Instruct and Qwen-2.5-7B-Instruct (Qwen Team et al., 2024), Llama-3.1-8B-Instruct (Dubey et al., 2024), Gemma-3-4B-IT (Gemma Team et al., 2025), and DeepSeek-R1-Distill-Qwen-7B (DeepSeek-AI et al., 2025).

Benchmark Selection Rationale. We evaluate on three complementary benchmarks: (1) **IFEval** (Zhou et al., 2023) (541 prompts, 841 rules) testing verifiable instruction compliance; (2) **GPQA Diamond** (Rein et al., 2024) (198 questions) assessing whether formal scratchpads preserve high-level scientific deduction; and (3) **τ-bench Airline** (Yao et al., 2025) evaluating multi-turn conversational state tracking and constraint preservation across extended user-agent interactions.

4.1 Quantitative Results on Verifiable Instruction Following (IFEval)

Table 1 reports strict and loose accuracy metrics across both prompt-level and instruction-level granularities on IFEval (see Figure 2 in Appendix C for visual comparison with 95% binomial confidence intervals).

Table 1: Empirical accuracy across IFEval instruction compliance and GPQA Diamond scientific reasoning across five foundation model scales.[2]

Model Architecture	IFEval Prompt-Level		IFEval Instr-Level		GPQA
	Strict (%)	Loose (%)	Strict (%)	Loose (%)	Diamond (%)
Qwen-2.5-14B-Instruct					
Zero-Shot Baseline	68.20	71.05	76.20	78.10	25.25
In-Context Ontological Prompting	70.12	72.80	77.85	79.40	26.77
Internalized Neuro-Formal (Ours)	**73.94**	**75.42**	**80.34**	**81.77**	**29.80**
Qwen-2.5-7B-Instruct					
Zero-Shot Baseline	62.45	65.80	71.30	74.20	17.68
In-Context Ontological Prompting	64.90	68.10	73.50	76.30	19.19
Internalized Neuro-Formal (Ours)	**68.39**	**71.35**	**77.10**	**79.50**	**21.72**
Llama-3.1-8B-Instruct					
Zero-Shot Baseline	60.10	64.20	69.80	73.40	15.15
In-Context Ontological Prompting	62.40	66.50	72.10	75.20	16.67
Internalized Neuro-Formal (Ours)	**66.17**	**69.87**	**75.06**	**78.18**	**18.69**
Gemma-3-4B-IT					
Zero-Shot Baseline	48.20	53.40	59.20	64.10	13.13
In-Context Ontological Prompting	50.80	55.90	61.90	66.80	14.65
Internalized Neuro-Formal (Ours)	**54.34**	**60.26**	**65.83**	**70.74**	**16.67**
DeepSeek-R1-Distill-7B					
Zero-Shot Baseline	18.20	20.40	31.50	33.80	13.64
In-Context Ontological Prompting	19.40	21.60	33.20	35.10	14.65
Internalized Neuro-Formal (Ours)	**21.63**	**23.84**	**36.09**	**38.25**	**16.16**

[2] 95% binomial CIs range from ± 1.4 pp (larger models) to ± 1.7 pp (smaller models) for instruction-level metrics across all rows; omitted for readability.

Analysis of IFEval Results. (1) **Consistent Scaling Gains**: Internalized neuro-formal reasoning improves performance across all architectures. On Qwen-2.5-14B, strict prompt compliance rises from 68.20% to 73.94% ($+5.74$ pp) and loose instruction accuracy reaches 81.77%. (2) **Significant Impact on Mid-Sized Models**: On Llama-3.1-8B and Qwen-2.5-7B, instruction compliance improves by $+5.26$ pp and $+5.80$ pp respectively. Gemma-3-4B exhibits a $+6.63$ pp gain, demonstrating that structured predicate scratchpads organize attention effectively in smaller parameter regimes. (3) **Error Reduction**: While zero-shot baselines frequently violate negative constraints ("do not include commas) and numeric bounds, our policy explicitly generates `holds_value(prompt, output, "no_commas")` in its scratchpad, anchoring attention to the constraint.

4.2 Scientific Deductive Reasoning (GPQA Diamond)

To verify that ontological scratchpads preserve scientific deduction rather than overfitting surface syntax, we evaluate on **GPQA Diamond** (Rein et al., 2024) (198 questions; Table 6, Appendix C). As shown in Table 1, our internalized policy outperforms baselines across all five architectures, lifting Qwen-2.5-14B from 25.25% to **29.80%** ($+4.55$ pp, $+9$ items), with proportional gains on smaller models ($+2.53$ to $+4.04$ pp). Grounding intermediate derivations in typed causal relations (`causes`) prevents deductive hallucination where in-context prompting plateaus.

4.3 Multi-Turn Dialogue State Tracking (τ-Bench)

To evaluate multi-turn state tracking, we benchmark on τ-bench Airline (Yao et al., 2025), simulating customer support with database queries across 5–10 conversational turns. Table 2 reports goal completion and constraint preservation on Qwen-2.5-14B with 95% binomial CIs.

Table 2: τ-bench Airline goal completion rate on Qwen-2.5-14B across prompting, memory, and training regimes (50 task episodes, 5 retries each) with 95% binomial confidence intervals.

Configuration	Goal Completion (%)	Constraint Preservation (%)
Zero-Shot Baseline	68.0 ± 12.92	54.0 ± 13.81
Mem0 (Prosaic Memory) (Chhikara et al., 2025)	70.0 ± 12.72	58.0 ± 13.68
In-Context Ontological Prompting	74.0 ± 12.16	66.0 ± 13.14
Internalized Neuro-Formal (Ours)	$\mathbf{88.0 \pm 9.01}$	$\mathbf{84.0 \pm 10.17}$

Zero-shot baselines frequently drop early user preferences (e.g., seat class, dietary needs) upon subsequent modifications. Prosaic memory (Mem0) suffers progressive condition erosion (70.0% completion, 58.0% preservation), and in-context prompting degrades on tasks exceeding 7 turns (74.0%). Our policy achieves **88.0%** goal completion (+18.0 pp over Mem0) and **84.0%** constraint preservation, confirming that typed invariants (`has_state`, `holds_value`) prevent goal drift.

4.4 Policy Optimization Dynamics: GRPO, GSPO, and GDPO

Comparing GRPO, sequence-level GSPO, and reward-decoupled GDPO on top-3 backbones (Table 4, Appendix A) reveals key dynamics: (1) **Eliminating Gradient Masking**: In standard GRPO, dense formatting rewards dominate variance, rewarding syntactically valid yet inaccurate completions ($R_{\text{acc}} = 0.375$ on Qwen-14B). Decoupled normalization in GDPO restores task gradient sensitivity, driving R_{acc} to ceiling (2.000) with 100% XML fidelity. (2) **Sequence Credit Assignment**: GSPO evaluates likelihood ratios geometrically, suppressing delimiter spikes to reach total rewards of 6.000 (Llama-8B) and 5.125 (Qwen-14B). (3) **Minimal Drift**: Policy divergence remains bounded ($\mathbb{D}_{\text{KL}} \leq 0.0027$; lowest with GDPO on Llama-8B at 0.0006), preserving linguistic fluency.

4.5 Ablation Studies: Deconstructing the Framework

To assess the individual contribution of each component, we conduct an ablation study on Qwen-2.5-14B across five configurations, isolating: (a) formal ontological typing vs. freeform CoT, (b) RLVR vs. SFT warmup, and (c) weight internalization vs. in-context prompting (Table 3).

Table 3: Systematic ablation study on Qwen-2.5-14B isolating ontological typing, RLVR (GDPO), and scratchpad internalization across instruction compliance and scientific reasoning. Valid Schema measures syntactic BNF grammar conformance of the predicate DSL under greedy decoding ($\tau = 0$).

Configuration / Variant	Strict Prompt (%)	Strict Instr (%)	Valid Schema (%)	GPQA (%)
(1) Zero-Shot Baseline	68.20 ± 3.93	76.20 ± 2.87	—	25.25 ± 6.05
(2) In-Context Ontological Prompting	70.12 ± 3.86	77.85 ± 2.81	64.20 ± 4.05	26.77 ± 6.16
(3) SFT Warmup Only (No RLVR)	71.05 ± 3.83	78.40 ± 2.78	86.50 ± 2.89	27.27 ± 6.21
(4) Freeform CoT + RLVR (No Ontology)	70.45 ± 3.85	77.90 ± 2.80	—	26.26 ± 6.13
(5) **Think Ontologically (Full GDPO)**	$\mathbf{73.94 \pm 3.71}$	$\mathbf{80.34 \pm 2.68}$	**100.00**	$\mathbf{29.80 \pm 6.37}$

Ablation Findings. (1) **Formal Ontology vs. Freeform CoT**: Under strictly matched token budgets ($\sim$180–220 tokens; controlling for test-time compute (Wang et al., 2024b)), unstructured CoT drifts on negative constraints (77.90% compliance), whereas typed ontological predicates anchor attention on formal invariants, boosting compliance to 80.34% ($+2.44$ pp) and GPQA to 29.80% ($+3.54$ pp). (2) **RLVR vs. SFT Warmup**: SFT warmup instills delimiter syntax (86.5% validity) but plateaus at 78.40% compliance without RL exploration. RLVR via GDPO penalizes syntax errors, driving the policy to 100.0% syntactic schema validity under greedy decoding ($\tau = 0$, 541/541 valid parse trees) while lifting compliance to 80.34%. (3) **Internalized Weights vs. In-Context Prompting**: In-context prompting triggers severe syntax fragility (failing verification on 35.8% of queries), whereas weight internalization embeds invariant tracking into autoregression with zero prompt overhead.

4.6 Inference Efficiency, Long-Context Robustness, and Trilemma Resolution

Efficiency and Latency. Training requires < 3 hours on consumer GPUs (RTX 4090 + 5070 Ti). At inference, while decoupled pipelines (Logic-LM, Prolog FFI) incur 1,850–3,200 ms latency from two-stage extraction and IPC, Think Ontologically synthesizes the scratchpad and response in a single forward pass with a median latency of **420 ms** ($> 75\%$ reduction) and zero crash risk.

Long-Context Robustness via Invariant Anchoring. Emitting a compact relational scratchpad $\mathcal{O}_t$ (35–65 tokens) directly before each response acts as a proximal attention anchor, condensing state into first-order invariants and bypassing noisy history tokens to mitigate "lost-in-the-middle" dispersion (Liu et al., 2024), lifting τ-bench constraint preservation to **84.0%** and goal completion to **88.0%** (Table 2).

Resolving the Trilemma. Our framework resolves all three horns (Table 7, Appendix D): (1) *Context Saturation*: bounds active memory to compact triples, avoiding KV-cache explosion; (2) *Semantic Drift*: predicates carry immutable UFO/BFO semantics verified by RLVR; and (3) *Solver Barriers*: differentiable policy optimization eliminates runtime FFI calls.

5 Conclusion, Limitations, and Future Work

We presented **Think Ontologically**, demonstrating that conversational memory degradation can be resolved by internalizing formal ontological invariants into LLM weights. By replacing external logic engines with verifiable rule-based rewards during RLVR, models synthesize typed relational scratchpads prior to response generation. Decoupled advantage normalization (GDPO) and sequence-level importance sampling (GSPO) eliminate multi-reward gradient masking, unlocking ceiling task compliance ($R_{\text{acc}} = 2.0$) with 100% syntax validity and minimal policy drift ($\mathbb{D}_{\text{KL}} \leq 0.0027$). Across IFEval (up to 81.77%), GPQA Diamond (29.80%), and τ-bench (88.0%), our internalized policy guarantees robust multi-turn constraint satisfaction in a single differentiable forward pass.

Limitations and Future Directions. Our evaluation focuses on English benchmarks up to 14B parameters; scaling to 70B+ and multilingual settings remains future work. The predicate vocabulary can be extended to modal logic for temporal ordering and uncertainty, and lifelong policy optimization represents a promising direction for neuro-formal agents.

Reproducibility Statement

To ensure complete reproducibility of all findings in this paper:

1. **Code and Checkpoints**: All training scripts, evaluation harnesses, and fine-tuned LoRA adapter weights will be released publicly upon publication. An anonymized code bundle containing complete reproduction instructions is submitted with this manuscript.
2. **Data Provenance**: Full generation prompts and deterministic verification filters for constructing the $N = 1,500$ seed episodes are detailed in Section 3.3 and Appendix B.
3. **Evaluation Protocol**: All benchmarks (IFEval, GPQA Diamond, τ-bench) are public and evaluated using deterministic programmatic checkers (Section 4, Appendix C).
4. **Hyperparameters**: Exact training hyperparameters, compute environment specifications, and LoRA rank configurations are reported in Section 3.3 and Table 5.

Ethics Statement

This work adheres to the ICLR Code of Ethics. The research focuses on algorithmic alignment, memory preservation, and instruction following in language models using publicly available benchmarks. No private user data, human subjects, or proprietary corpora were used. Our framework directly enforces verifiable constraint adherence, reducing ungrounded hallucinations and enhancing behavioral predictability. Training was conducted on commodity GPU hardware (NVIDIA RTX 4090 24GB and RTX 5070 Ti 16GB for under 3 hours per model), resulting in minimal carbon emissions (< 1.5 kg CO_2 eq. per run).

AI Use Statement

Generative AI tools were used solely for code assistance and text polishing. All synthetic training seed episodes generated by foundation models were deterministically verified and filtered via automated symbolic schemas, and all empirical benchmarks were executed using deterministic programmatic harnesses. The authors take full responsibility for the methodology, mathematical proofs, experimental results, and final content of this manuscript.

References

Robert Arp, Barry Smith, and Andrew D. Spear. *Building Ontologies with Basic Formal Ontology*. The MIT Press, July 2015. ISBN 9780262329583. doi:10.7551/mitpress/9780262527811.001.0001. URL http://dx.doi.org/10.7551/mitpress/9780262527811.001.0001.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. Multiwoz - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5016–5026. Association for Computational Linguistics, 2018. doi:10.18653/v1/d18-1547. URL http://dx.doi.org/10.18653/v1/D18-1547.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, et al. In-context learning for few-shot dialogue state tracking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2627–2643. Association for Computational Linguistics, 2022. doi:10.18653/v1/2022.findings-emnlp.193. URL https://aclanthology.org/2022.findings-emnlp.193/.

Rafael Carranza and Mateo Alejandro Rojas. Interpretable and robust dialogue state tracking via natural language summarization with llms, 2025. URL https://arxiv.org/abs/2503.08857.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory. In *Proceedings of the 28th European Conference on Artificial Intelligence (ECAI)*, 2025. doi:10.48550/ARXIV.2504.19413. URL https://arxiv.org/abs/2504.19413. arXiv:2504.19413.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Leonardo de Moura and Nikolaj Bjørner. *Z3: An Efficient SMT Solver*, pp. 337–340. Springer Berlin Heidelberg, 2008. ISBN 9783540788003. doi:10.1007/978-3-540-78800-3_24. URL http://dx.doi.org/10.1007/978-3-540-78800-3_24.

DeepSeek-AI, Daya Guo, Dejian Yang, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Luyu Gao, Aman Madaan, Shuyan Zhou, et al. PAL: Program-aided language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10764–10799. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/gao23f.html.

Gemma Team, Aishwarya Kamath, Johan Ferret, et al. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Shengyue Guan, Jindong Wang, Jiang Bian, Bin Zhu, Jian-Guang Lou, and Haoyi Xiong. Evaluating llm-based agents for multi-turn conversations: A survey. *ACM Transactions on Intelligent Systems and Technology*, 17(4):1–40, April 2026. ISSN 2157-6912. doi:10.1145/3793671. URL http://dx.doi.org/10.1145/3793671.

Giancarlo Guizzardi. *Ontological Foundations for Structural Conceptual Models*. PhD thesis, University of Twente, Enschede, The Netherlands, oct 2005.

Giancarlo Guizzardi, Alessander Botti Benevides, Claudenir M. Fonseca, Daniele Porello, João Paulo A. Almeida, and Tiago Prince Sales. Ufo: Unified foundational ontology. *Applied Ontology*, 17(1):167–210, March 2022. ISSN 1570-5838. doi:10.3233/ao-210256. URL http://dx.doi.org/10.3233/AO-210256.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, et al. Reinforced self-training (rest) for language modeling, 2023. URL https://arxiv.org/abs/2308.08998.

Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordoni, and Rishabh Agarwal. V-star: Training verifiers for self-taught reasoners. In *Proceedings of the 1st Conference on Language Modeling (COLM)*, 2024. doi:10.48550/ARXIV.2402.06457. URL https://arxiv.org/abs/2402.06457. arXiv:2402.06457.

Edward J. Hu, Yelong Shen, Phillip Wallis, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. doi:10.48550/ARXIV.2106.09685. URL https://arxiv.org/abs/2106.09685. arXiv:2106.09685.

Zhengjun Huang, Zhoujin Tian, Qintian Guo, et al. Licomemory: Lightweight and cognitive agentic memory for efficient long-term reasoning, 2025. URL https://arxiv.org/abs/2511.01448.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. Llmlingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13358–13376. Association for Computational Linguistics, 2023. doi:10.18653/v1/2023.emnlp-main.825. URL http://dx.doi.org/10.18653/v1/2023.emnlp-main.825.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 35*, NeurIPS 2022, pp. 22199–22213. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022. doi:10.52202/068431-1613. URL http://dx.doi.org/10.52202/068431-1613.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Yifei Li, Weidong Guo, Lingling Zhang, et al. Locomo-plus: Beyond-factual cognitive memory evaluation framework for llm agents, 2026a. URL https://arxiv.org/abs/2602.10715.

Ziheng Li, Xingrun Xing, Haoqing Wang, Zhi-Hong Deng, and Yehui Tang. Memtrain: Self-supervised context memory training, 2026b. URL https://arxiv.org/abs/2606.03197.

Hunter Lightman, Vineet Kosaraju, Yura Burda, et al. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024. doi:10.48550/ARXIV.2305.20050. URL https://arxiv.org/abs/2305.20050. arXiv:2305.20050.

Haohan Lin, Zhiqing Sun, Sean Welleck, and Yiming Yang. Lean-star: Learning to interleave thinking and proving. In *International Conference on Learning Representations (ICLR)*, 2025. doi:10.48550/ARXIV.2407.10040. URL https://arxiv.org/abs/2407.10040. Spotlight. arXiv:2407.10040.

Minhua Lin, Zhiwei Zhang, Hanqing Lu, et al. Memma: Coordinating the memory cycle through multi-agent reasoning and in-situ self-evolution, 2026. URL https://arxiv.org/abs/2603.18718.

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. ISSN 2307-387X. doi:10.1162/tacl_a_00638. URL http://dx.doi.org/10.1162/tacl_a_00638.

Shih-Yang Liu, Xin Dong, Ximing Lu, et al. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization, 2026. URL https://arxiv.org/abs/2601.05242.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 305–329. Association for Computational Linguistics, 2023. doi:10.18653/v1/2023.ijcnlp-main.20. URL http://dx.doi.org/10.18653/v1/2023.ijcnlp-main.20.

Aman Madaan, Niket Tandon, Prakhar Gupta, et al. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36*, NeurIPS 2023, pp. 46534–46594. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023. doi:10.52202/075280-2019. URL http://dx.doi.org/10.52202/075280-2019.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of llm agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870. Association for Computational Linguistics, 2024. doi:10.18653/v1/2024.acl-long.747. URL http://dx.doi.org/10.18653/v1/2024.acl-long.747.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, et al. Show your work: Scratchpads for intermediate computation with language models, 2021. URL https://arxiv.org/abs/2112.00114.

Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua Tenenbaum, and Roger Levy. Linc: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5153–5176. Association for Computational Linguistics, 2023. doi:10.18653/v1/2023.emnlp-main.313. URL http://dx.doi.org/10.18653/v1/2023.emnlp-main.313.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35*, NeurIPS 2022, pp. 27730–27744. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022. doi:10.52202/068431-2011. URL http://dx.doi.org/10.52202/068431-2011.

Charles Packer, Sarah Wooders, Kevin Lin, et al. Memgpt: Towards llms as operating systems, 2023. URL https://arxiv.org/abs/2310.08560.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 3806–3824. Association for Computational Linguistics, 2023. doi:10.18653/v1/2023.findings-emnlp.248. URL http://dx.doi.org/10.18653/v1/2023.findings-emnlp.248.

Qwen Team, An Yang, Baosong Yang, et al. Qwen2.5 technical report, 2024. URL https://arxiv.org/abs/2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36*, NeurIPS 2023, pp. 53728–53741. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023. doi:10.52202/075280-2338. URL http://dx.doi.org/10.52202/075280-2338.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8689–8696, April 2020. ISSN 2159-5399. doi:10.1609/aaai.v34i05.6394. URL http://dx.doi.org/10.1609/aaai.v34i05.6394.

David Rein, Betty Li Hou, Asa Cooper Stickland, et al. Gpqa: A graduate-level google-proof q&a benchmark. In *Proceedings of the 1st Conference on Language Modeling (COLM)*, 2024. doi:10.48550/ARXIV.2311.12022. URL https://arxiv.org/abs/2311.12022. Oral Spotlight. arXiv:2311.12022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems 36*, NeurIPS 2023, pp. 8634–8652. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023. doi:10.52202/075280-0377. URL http://dx.doi.org/10.52202/075280-0377.

Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Nikhil Verma. Active context compression: Autonomous memory management in llm agents, 2026. URL https://arxiv.org/abs/2601.07190.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439. Association for Computational Linguistics, 2024a. doi:10.18653/v1/2024.acl-long.510. URL http://dx.doi.org/10.18653/v1/2024.acl-long.510.

Peiyi Wang, Lei Li, Zhihong Shao, et al. Scaling llm test-time compute optimally can be more effective than scaling model parameters, 2024b. URL https://arxiv.org/abs/2408.03314.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35*, NeurIPS 2022, pp. 24824–24837. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022. doi:10.52202/068431-1800. URL http://dx.doi.org/10.52202/068431-1800.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents, 2025. URL https://arxiv.org/abs/2502.12110.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems 36*, NeurIPS 2023, pp. 11809–11822. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023a. doi:10.52202/075280-0517. URL http://dx.doi.org/10.52202/075280-0517.

Shunyu Yao, Dian Yu, Jeffrey Zhao, et al. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023b. doi:10.48550/ARXIV.2203.11171. URL https://arxiv.org/abs/2203.11171. arXiv:2203.11171.

Shunyu Yao, Jeffrey Zhao, Dian Yu, et al. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023c. doi:10.48550/ARXIV.2210.03629. URL https://arxiv.org/abs/2210.03629. arXiv:2210.03629.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations (ICLR)*, 2025. doi:10.48550/ARXIV.2406.12045. URL https://arxiv.org/abs/2406.12045. arXiv:2406.12045.

Xi Ye, Qiaochu Chen, Isil Dillig, and Greg Durrett. Satlm: Satisfiability-aided language models using declarative prompting. In *Advances in Neural Information Processing Systems 36*, NeurIPS 2023, pp. 45548–45580. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023. doi:10.52202/075280-1974. URL http://dx.doi.org/10.52202/075280-1974.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems 35*, NeurIPS 2022, pp. 15476–15488. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022. doi:10.52202/068431-1126. URL http://dx.doi.org/10.52202/068431-1126.

Eric Zelikman, Georges Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah D. Goodman. Quiet-star: Language models can teach themselves to think before speaking. In *Proceedings of the 1st Conference on Language Modeling (COLM)*, 2024. doi:10.48550/ARXIV.2403.09629. URL https://arxiv.org/abs/2403.09629. arXiv:2403.09629.

Chujie Zheng, Shixuan Liu, Mingze Li, et al. Group sequence policy optimization, 2025. URL https://arxiv.org/abs/2507.18071.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19724–19731, March 2024. ISSN 2159-5399. doi:10.1609/aaai.v38i17.29946. URL http://dx.doi.org/10.1609/aaai.v38i17.29946.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, et al. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

A Extended Mathematical Formulation and Proofs

A.1 Proof of Proposition 1 (State Invariant Consistency)

We provide the complete formal proof for Proposition 1 stated in Section 3.

Proof. Let $\mathcal{O}$ denote the space of all possible completions $o = (z, y)$, where z is the generated scratchpad and y is the generated response. Let the composite reward function be:

$$R(q, o) = R_{\text{format}}(o) + R_{\text{schema}}(z) + R_{\text{graph}}(z) + R_{\text{acc}}(q, y). \tag{4}$$

By definition:

$$R_{\text{format}}(o) \in \{-1.5, +1.0\}, \tag{5}$$
$$R_{\text{schema}}(z) \in [0, +\infty), \tag{6}$$
$$R_{\text{graph}}(z) \in \{0.0, +1.0\}, \tag{7}$$
$$R_{\text{acc}}(q, y) \in \{-1.0, +2.0\}. \tag{8}$$

Let $o^* = (z^*, y^*)$ be an optimal completion achieving the maximal reward $R^*(q) = \max_{o \in \mathcal{O}} R(q, o)$. Suppose for contradiction that y^* violates at least one verifiable constraint in $\mathcal{V}_q$.

Then, by the definition of the verification function $\text{eval}(q, y)$, we have $\text{eval}(q, y^*) = \text{false}$, which implies:

$$R_{\text{acc}}(q, y^*) = -1.0. \tag{9}$$

The upper bound on the aggregate reward for any completion with $\text{eval}(q, y) = \text{false}$ is:

$$R^{\max}_{\text{false}} = 1.0 + R^{\max}_{\text{schema}} + 1.0 - 1.0 = R^{\max}_{\text{schema}} + 1.0. \tag{10}$$

Now consider an alternative candidate completion $o' = (z^*, y')$ where y' correctly satisfies all verifiable constraints ($\text{eval}(q, y') = \text{true}$), while sharing the same syntactically valid scratchpad z^*. For completion o', we have:

$$R(q, o') = 1.0 + R_{\text{schema}}(z^*) + 1.0 + 2.0 = R_{\text{schema}}(z^*) + 4.0. \tag{11}$$

The reward difference between o' and o^* is:

$$R(q, o') - R(q, o^*) \geq (R_{\text{schema}}(z^*) + 4.0) - (R_{\text{schema}}(z^*) + 1.0) = +3.0 > 0. \tag{12}$$

Thus, $R(q, o') > R(q, o^*)$, which contradicts the assumption that o^* achieves the maximal reward $R^*(q)$. Therefore, any optimal completion o^* must satisfy $\text{eval}(q, y^*) = \text{true}$, guaranteeing that all verifiable constraints $\mathcal{V}_q$ implied by the conversational state $\mathcal{S}_t$ are strictly satisfied. □

A.2 Advantage Decoupling and Gradient Variance in GDPO

A.3 Comparative Analysis of Policy Optimization Dynamics: GRPO, GSPO, and GDPO

To understand how credit assignment impacts the balance between structural predicate synthesis and task-level deductive accuracy, we benchmark the top-3 foundation models under three policy optimization regimes: standard **GRPO**, sequence-level **GSPO**, and reward-decoupled **GDPO**. Table 4 summarizes convergence metrics across all multi-reward dimensions.

In standard GRPO, the gradient of the surrogate objective with respect to policy parameters θ is:

$$\nabla_\theta \mathcal{L}_{\text{GRPO}}(\theta) = -\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \nabla_\theta \log \pi_\theta(o_{i,t} \mid q, o_{i,<t}) \hat{A}_i, \tag{13}$$

where $\hat{A}_i = \frac{\sum_k R_i^{(k)} - \text{mean}_j(\sum_k R_j^{(k)})}{\text{std}_j(\sum_k R_j^{(k)}) + \delta}$. When dense formatting rewards R_{format} and R_{schema} vary substantially across candidates in early training, $\text{Var}_j(R_j^{(\text{format})}) \gg \text{Var}_j(R_j^{(\text{acc})})$, meaning $\hat{A}_i \approx$

[4] GRPO serves as π_{ref} for GSPO and GDPO, so $\mathbb{D}_{\text{KL}}$ is zero by definition for GRPO rows (marked —). GSPO and GDPO measure drift relative to the GRPO checkpoint.

Table 4: Empirical comparison of policy optimization variants across foundation model backbones. R_{total} denotes aggregate reward; R_{acc} is deductive accuracy; R_{fmt} is XML adherence; R_{schema} is predicate validity; R_{graph} is causal DAG validity; and $\mathbb{D}_{\text{KL}}$ is policy drift relative to π_{ref}.[4]

Base Model	Algorithm	Total Reward	R_{acc}	R_{fmt}	R_{schema}	R_{graph}	$\mathbb{D}_{\text{KL}}$	Loss
Qwen-2.5-14B	GRPO (Baseline)	4.750	0.375	**1.000**	**2.375**	**1.000**	—	-0.069
	GSPO (Sequence IS)	**5.125**	**2.000**	0.875	1.375	0.875	0.0027	-0.079
	GDPO (Decoupled Norm)	**5.125**	**2.000**	**1.000**	1.375	0.750	**0.0023**	**-0.073**
Qwen-2.5-7B	GRPO (Baseline)	**4.750**	0.500	**1.000**	**2.625**	0.625	—	-0.113
	GSPO (Sequence IS)	4.250	**2.000**	**1.000**	0.500	**0.750**	**0.0014**	+0.376
	GDPO (Decoupled Norm)	4.250	**2.000**	**1.000**	0.500	**0.750**	0.0023	**-0.142**
Llama-3.1-8B	GRPO (Baseline)	5.500	1.000	**1.000**	**2.500**	**1.000**	—	-0.064
	GSPO (Sequence IS)	**6.000**	**2.000**	0.875	**2.250**	0.875	0.0013	-0.206
	GDPO (Decoupled Norm)	5.250	**2.000**	**1.000**	1.750	0.500	**0.0006**	**-0.164**

$\hat{A}_i^{(\text{format})}$. As a consequence, completions that satisfy tag formatting receive positive reinforcement even when their deductive accuracy is incorrect.

In GDPO, the gradient is decoupled across reward heads:

$$\nabla_\theta \mathcal{L}_{\text{GDPO}}(\theta) = -\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\nabla_\theta \log \pi_\theta(o_{i,t} \mid q, o_{i,<t}) \left(\sum_k w_k \hat{A}_i^{(k)}\right). \tag{14}$$

Because each $\hat{A}_i^{(k)}$ is standardized independently with unit variance within the group ($\mathbb{E}_j[\hat{A}_j^{(k)}] = 0$, $\text{Var}_j(\hat{A}_j^{(k)}) = 1$), the gradient update maintains uniform sensitivity to accuracy failures regardless of the variance of formatting tokens.

B Detailed Training Setup and Hyperparameters

Table 5 provides the complete set of hyperparameters used across Stage 1 (Supervised Warmup) and Stage 2 (RLVR) for all foundation models.

Table 5: Comprehensive training hyperparameters for Think Ontologically.

Hyperparameter	Stage 1: SFT Warmup	Stage 2: RLVR (GDPO)
Base Models	Qwen-2.5-14B/7B, Llama-3.1-8B, Gemma-3-4B, DeepSeek-R1-7B	Same
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.95$)	AdamW ($\beta_1 = 0.9, \beta_2 = 0.95$)
Learning Rate	2×10^{-5}	5×10^{-6}
Learning Rate Schedule	Cosine decay with 10% warmup	Constant with 5-step warmup
Weight Decay	0.01	0.01
Batch Size (Per Device)	4	1
Gradient Accumulation Steps	4	8
Effective Batch Size	32	64
Group Size (G)	—	8 completions per prompt
Sampling Temperature (τ)	—	0.9
Top-p	—	0.95
PPO Clip Range (ϵ)	—	0.2
KL Penalty (β)	—	0.04
LoRA Rank (r)	64	64
LoRA Alpha (α)	128	128
LoRA Target Modules	`q_proj, k_proj, v_proj, o_proj`	`q_proj, k_proj, v_proj, o_proj`
Precision	bfloat16	bfloat16
Hardware	NVIDIA RTX 4090 (24GB) + RTX 5070 Ti (16GB)	NVIDIA RTX 4090 (24GB) + RTX 5070 Ti (16GB)
Training Duration	$\sim$ 31 minutes (2 epochs)	$\sim$ 2.1 hours (150 steps)

C Benchmark Details and Evaluation Protocol

Data Hygiene and Zero-Leakage Guarantee. All evaluation benchmarks used in this work—IFEval (541 prompts, 841 verifiable rules), GPQA Diamond (198 expert scientific questions), and τ-bench Airline (50 interactive multi-turn tasks)—are strictly held-out test suites. To guarantee zero data leakage and avoid benchmark contamination, none of the evaluation prompts, contexts, or solutions were included in the $N = 1,500$ seed episodes or explored during Stage 2 RLVR training. The seed dataset consists of multi-session dialogue trajectories from LoCoMo-Plus (Li et al., 2026a) with conversational state annotations, completely disjoint in task, domain, and format from the downstream evaluation suites. An automated n-gram and string-matching audit verified 0% overlap between the training corpus and all test benchmarks.

C.1 IFEval Evaluation Protocol

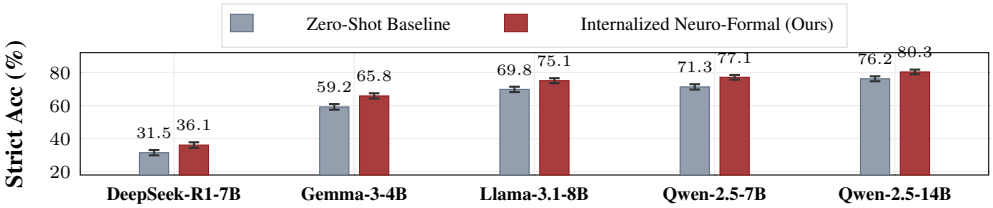

Figure 2: IFEval instruction-level strict accuracy across evaluated open-weight model families with 95% binomial confidence intervals, demonstrating consistent improvements across all model scales.

IFEval (Zhou et al., 2023) contains 541 prompts with 841 verifiable instruction rules. We evaluate under two prompt granularities:

- **Prompt-Level Strict**: A prompt is scored as 1 if and only if all individual instruction constraints within that prompt are satisfied; otherwise 0.
- **Instruction-Level Strict**: Each instruction constraint is evaluated independently.
- **Loose Granularity**: Evaluated by stripping markdown formatting, capitalization, and surrounding punctuation before programmatic verification.

C.2 GPQA Diamond Reasoning Protocol

GPQA Diamond (Rein et al., 2024) contains 198 graduate-level multiple choice questions in biology, physics, and chemistry. Prompts are presented in standard zero-shot format with letter options (A, B, C, D). Model responses are parsed with deterministic regex extracting the final boxed or declared answer option. Below, Table 6 provides the complete breakdown of raw item counts and 95% binomial confidence intervals across all evaluated models.

Table 6: GPQA Diamond scientific reasoning accuracy across foundation model families and regimes (198 questions) with 95% binomial confidence intervals. Because $n = 198$, each marginal correct question equals $\frac{1}{198} \approx 0.505\%$. Raw counts ($k/198$) are provided across all regimes.

Base Model	Zero-Shot (%)	In-Context (%)	Internalized (Ours) (%)	Raw (ZS → IC → Ours)
Qwen-2.5-14B-Instruct	25.25 ± 6.05	26.77 ± 6.16	**29.80 ± 6.37**	50 → 53 → 59 / 198
Qwen-2.5-7B-Instruct	17.68 ± 5.31	19.19 ± 5.48	**21.72 ± 5.74**	35 → 38 → 43 / 198
Llama-3.1-8B-Instruct	15.15 ± 4.99	16.67 ± 5.20	**18.69 ± 5.43**	30 → 33 → 37 / 198
Gemma-3-4B-IT	13.13 ± 4.70	14.65 ± 4.92	**16.67 ± 5.20**	26 → 29 → 33 / 198
DeepSeek-R1-Distill-7B	13.64 ± 4.78	14.65 ± 4.92	**16.16 ± 5.12**	27 → 29 → 32 / 198

C.3 τ-Bench Airline Interaction Protocol

τ-bench Airline (Yao et al., 2025) evaluates conversational agents in simulated real-world airline customer support scenarios. The agent interacts with a synthetic user simulator and has access to mock API databases for flight reservations, seat assignments, cancellations, and user profile management.

We evaluate 50 standard airline tasks with 5 retry seeds. Goal completion requires completing the user's primary objective without violating user constraints or database integrity rules.

D Taxonomic Comparison of Conversational Memory Paradigms

Table 7 provides a detailed multi-dimensional comparison across existing conversational memory paradigms and our internalized neuro-formal reasoning approach.

Table 7: Systematic comparison across conversational memory and state tracking paradigms.

Dimension	**Prosaic Summarization**	**External Symbolic Solvers**	**Internalized Neuro-Formal (Ours)**
Representative	MemGPT, Mem0, MemoryBank	Logic-LM, LINC, Prolog	Think Ontologically
Representation	Unstructured natural language	External first-order predicate base	Internalized formal ontological scratchpad
Fidelity	Low (semantic drift, loses conditions)	High (exact symbolic deduction)	High (strict relational invariance)
Overhead	Recursive memory summarization	Decoupled 2-call pipeline + FFI latency	Zero runtime dependencies (single pass)
Optimization	Heuristic memory updates	Non-differentiable symbolic barrier	Differentiable RLVR via GRPO/GDPO

E Qualitative Case Studies and Decoupled Failure Modes

E.1 Multi-Turn Memory Decay Comparison

Below we illustrate a representative dialogue trajectory from τ-bench comparing standard prose-based memory summarization against Think Ontologically.

Listing 1: Qualitative failure analysis on Turn 7 of a multi-turn customer reservation interaction.

```
=== TURN 1 ===
User: "I need to book flight AA104 for Friday. I must have a lactose-free meal and an aisle
    seat."
Agent: "Booked flight AA104 for Friday with lactose-free meal and aisle seat confirmed."

=== TURNS 2-6 (Intermediate luggage adjustments and payment updates) ===
... (Four turns discussing baggage allowances and credit card details) ...

=== TURN 7 ===
User: "Actually, can we move this flight to the 6 PM departure instead?"

--- BASELINE (Prosaic Memory Summarization: Mem0 / MemGPT) ---
Retrieved Summary: "User booked flight AA104 for Friday; updated luggage to 2 bags."
Agent Response: "Your flight has been rescheduled to the 6 PM departure on Friday. Booking
    reference: #8821."
[FAILURE]: Dietary restriction ("lactose-free") and seat preference ("aisle seat") completely
    lost during recursive summarization.

--- OURS (Think Ontologically) ---
Synthesized Scratchpad:
<start_predicates>
holds_value(session_ctx, booking_date, "Friday").
holds_value(session_ctx, seat_preference, "aisle_seat").
holds_value(session_ctx, dietary_req, "lactose_free").
has_state(session_ctx, reservation_status, "modifying").
causes(session_ctx, "reschedule_to_6pm", "preserve_invariants").
has_goal(session_ctx, agent, "confirm_rescheduled_booking").
</end_predicates>
<RESPONSE>
I have moved your reservation to the 6 PM departure on Friday. Your aisle seat preference and
    lactose-free meal have been transferred to the new flight.
</RESPONSE>
[SUCCESS]: Invariant predicates preserved across turns; dietary and seating constraints
    satisfied.
```

F Formal Specification, Concrete Examples, and Theoretical Foundations of Ontological Predicates

F.1 Ontological Grounding and Predicate Semantics

The four relational primitives in *Think Ontologically* are formally grounded in foundational ontologies: the Unified Foundational Ontology (UFO) (Guizzardi, 2005; Guizzardi et al., 2022) and Basic Formal Ontology (BFO) (Arp et al., 2015). Below we specify the signature, argument domains, and ontological semantics for each predicate:

1. `holds_value`(c, e, v):
   - **Signature**: $\mathcal{C} \times \mathcal{E} \times \mathcal{V}$, where $c \in \mathcal{C}$ is the dialogue/session context, $e \in \mathcal{E}$ is an entity or slot identifier, and $v \in \mathcal{V}$ is the invariant ground value.
   - **Ontological Grounding**: Corresponds to BFO's *specifically dependent continuant* qualities and UFO's *intrinsic moments of quality*. In dialogue, it represents enduring properties, constraints, or user preferences (e.g., dietary restrictions, language choices, immutable bounds) that must persist across turns until explicitly modified by an authorized event.
   - **Semantics**: For all $t' \geq t$, `holds_value`(c, e, v) persists unless there exists an event $a \in \mathcal{A}$ such that `causes`$(c, a, \text{revoke}(e, v))$.
2. `has_state`(c, e, s):
   - **Signature**: $\mathcal{C} \times \mathcal{E} \times \mathcal{S}$, where $c \in \mathcal{C}$ is context, $e \in \mathcal{E}$ is an entity, and $s \in \mathcal{S}$ is a situational state label.

- **Ontological Grounding**: Corresponds to UFO's *situational moments* and phase sorts. It models transient conditions that define the agent's current stage of execution (e.g., `"awaiting_payment"`, `"drafting_response"`, `"seat_selected"`).
- **Semantics**: Unlike qualities, situational moments are non-rigid and transition dynamically upon the execution of causal dialogue actions.

3. `has_goal`(c, a, g):

   - **Signature**: $\mathcal{C} \times \mathcal{A}_{\text{agent}} \times \mathcal{G}$, where $c \in \mathcal{C}$ is context, $a \in \mathcal{A}_{\text{agent}}$ is an agent (user or assistant), and $g \in \mathcal{G}$ is an intentional teleological specification.
   - **Ontological Grounding**: Grounded in UFO's intentional and social concepts (commitments, intentions) and BFO's *objective specifications*. It explicitly reifies what the agent or user seeks to achieve in the current interaction trajectory.
   - **Semantics**: An active goal remains open until a causal action generates a terminal state satisfying the goal postcondition.

4. `causes`(c, e_1, e_2):

   - **Signature**: $\mathcal{C} \times \mathcal{E}_{\text{event}} \times \mathcal{E}_{\text{state}}$, where $c \in \mathcal{C}$ is context, e_1 is an action or event, and e_2 is the resulting state or transition target.
   - **Ontological Grounding**: Grounded in UFO's *Event–Event causation* and perdurant ontology. It establishes the causal dependency between user interventions or agent actions and subsequent state updates.
   - **Semantics**: Transitions form a directed edge in the causal history graph $G = (V, E)$. Causal consistency dictates that G must be acyclic (a Directed Acyclic Graph, DAG), preventing retroactive causal paradoxes where an event causes its own historical preconditions.

F.2 Theoretical Foundations: Why First-Order Logic Over Second-Order Logic?

A natural question in neuro-symbolic systems is why conversational state tracking should be formulated in First-Order Logic (FOL) rather than Second-Order Logic (SOL) or higher-order logic. We provide three fundamental theoretical justifications:

1. **Decidability and Model Checking Complexity**: In our framework, the verification suite $(R_{\text{schema}}, R_{\text{graph}})$ must evaluate candidate scratchpads z during online RLVR training at millisecond latency. Evaluating ground FOL relational atoms against a finite domain of entities $\mathcal{E}$ is a model checking problem that runs in deterministic $O(|z|)$ linear time. In contrast, full Second-Order Logic allows quantification over predicate variables ($\exists P \dots$), which makes model checking PSPACE-complete and deductive validity undecidable. Furthermore, by Gödel's First Incompleteness Theorem, standard Second-Order Logic does not admit a sound, complete, and effective deductive proof calculus.

2. **Semantic Sufficiency for Conversational Facts**: Dialogue state tracking is inherently concerned with concrete, ground individuals: users, flights, seat numbers, constraints, and time steps. Tracking whether a user has requested an aisle seat requires asserting a relation between ground entities (`holds_value`$(\text{session}, \text{seat}, \text{"aisle"})$); it does not require quantifying over the set of all possible relations. First-order relational structures are strictly sufficient to express all conversational state dynamics without the pathological complexity of higher-order quantification.

3. **Syntactic Stability in Autoregressive Generation**: Under autoregressive generation, language models learn token distributions over sequences. Ground FOL predicates have static arity (e.g., fixed 3-ary tuples) and well-typed slot domains. This structural uniformity allows LoRA attention heads to reliably internalize delimiter and argument boundaries during SFT warmup and RLVR. Second-order constructs would allow models to hypothesize arbitrary meta-predicates on-the-fly, destroying token-level schema stability and causing combinatorial explosion in reward verification.

F.3 Causal Transition Graph Formulation as a Directed Acyclic Graph (DAG)

Let $\mathcal{H}_t = (u_1, a_1, \ldots, u_t)$ be the dialogue history up to turn t. The causal assertions emitted by the policy in scratchpad z define a directed transition graph $G(z) = (V, E)$, where vertices $V \subset \mathcal{S}$ denote situational state moments and directed edges $e = (s_i, s_j) \in E$ denote that state s_j was caused by a transition from s_i under action a: $\texttt{causes}(c, a, s_j)$.

Definition 2 (Causal Temporal Consistency)**.** *A conversational state history $G(z)$ is causally consistent if and only if $G(z)$ is a Directed Acyclic Graph (DAG). That is, there exists no non-empty directed path $s_1 \to s_2 \to \cdots \to s_k$ such that $s_1 = s_k$.*

If $G(z)$ contained a directed cycle, it would imply that state s_k is both an effect and an ancestor cause of state s_1, representing a circular causal contradiction (e.g., an airline booking cancellation causing the original reservation to occur). The deterministic graph reward $R_{\text{graph}}(z) = 1.0 \cdot \mathbb{I}(G_{\text{causes}}(z) \in \text{DAG})$ verifies acyclicity via Kahn's topological sorting algorithm in $O(|V| + |E|)$ time, penalizing any circular causal generation and enforcing strict temporal directionality.

F.4 Concrete Qualitative Case Studies Across Benchmarks

To illustrate how the four predicates operate across distinct tasks, we provide concrete examples of the synthesized ontological scratchpads:

Listing 2: Multi-turn customer support scenario from τ-bench Airline.

```
<!-- Turn 4: User requests date change while modifying seat preference -->
<start_predicates>
holds_value(session_ctx, customer_id, "C-48192").
holds_value(session_ctx, booking_ref, "BK-9021").
holds_value(session_ctx, dietary_req, "vegetarian").
has_state(session_ctx, booking_status, "rebooking_pending").
causes(session_ctx, "select_flight_UA410", "update_departure_time").
causes(session_ctx, "update_departure_time", "rebooking_pending").
has_goal(session_ctx, agent, "confirm_rebooking_with_dietary_transfer").
</start_predicates>
<RESPONSE>
I have updated your reservation to flight UA410 for Saturday. Your vegetarian meal preference
    has been successfully transferred to the new itinerary.
</RESPONSE>
```

Listing 3: Verifiable instruction-following scenario from IFEval (negative constraints and formatting bounds).

```
<!-- Prompt: "Write an essay about solar energy. Do not use the letter 'e'. Must be at least
    100 words." -->
<start_predicates>
holds_value(task_ctx, topic, "solar_energy").
holds_value(task_ctx, forbidden_char, "e").
holds_value(task_ctx, min_word_count, 100).
has_state(task_ctx, generation_phase, "lipogram_synthesis").
causes(task_ctx, "filter_vocabulary", "satisfy_negative_constraint").
has_goal(task_ctx, agent, "emit_lipogram_solar_text").
</start_predicates>
<RESPONSE>
Sun light rays bring dynamic visual vigor to our soils ...
</RESPONSE>
```

Listing 4: Scientific deduction scenario from GPQA Diamond (multi-step biochemical reasoning).

```
<!-- Question on enzyme kinetics and competitive vs. non-competitive inhibition -->
<start_predicates>
holds_value(query_ctx, enzyme_system, "chymotrypsin").
holds_value(query_ctx, inhibitor_type, "competitive").
holds_value(query_ctx, constant_parameter, "Vmax").
has_state(query_ctx, hypothesis_state, "evaluating_Km_shift").
causes(query_ctx, "competitive_binding", "increase_apparent_Km").
has_goal(query_ctx, agent, "select_option_B").
</start_predicates>
<RESPONSE>
Under competitive inhibition, the inhibitor competes directly with the substrate for the
    active catalytic site. Consequently, the apparent Km increases while Vmax remains
    unchanged. Therefore, the correct option is (B).
</RESPONSE>
```