
Hybrid Latent Reasoning via Reinforcement Learning

**Zhenrui Yue¹, Bowen Jin¹, Huimin Zeng¹, Honglei Zhuang², Zhen Qin², Jinsung Yoon²,
Lanyu Shang³, Jiawei Han¹, Dong Wang¹**

¹University of Illinois Urbana-Champaign, ²Google, ³LMU

{zhenrui3,bowenj4,huiminz3,lshang3,hanj,dwang24}@illinois.edu,

{hlz,zhenqin,jinsungyoon}@google.com, lanyu.shang@lmu.edu

Background: From Discrete CoT to Latent Reasoning

Paradigm Shift

- ▶ **Traditional CoT:** Relies on *discrete* decoding and sampling (visible tokens).
- ▶ **Latent Reasoning:** Enables LLMs to reason internally using *continuous hidden representations*.

Limitations

- ▶ **Dependence on CoT Data:** Methods like CODI rely on distilling discrete CoT traces.
- ▶ **High Training Costs:** Requires multi-stage training.
- ▶ **The Incompatibility Issue:** Output hidden states \neq Input embeddings. Direct feeding causes repetition and incoherence.

Proposed Solution: HRPO

Core Idea: Unify policy learning with latent reasoning via Reinforcement Learning (RL).

- ▶ **Problem Solved:** Mitigates the discrepancy between hidden states and embeddings.
- ▶ **Mechanism:**
 - ▶ **Start:** Prioritizes sampled **Token Embeddings** (Preserves generative ability).
 - ▶ **During Training:** Gradually incorporates **Continuous Hidden States** (Enables internal reasoning).
- ▶ **Optimization (RL):** Uses simple **outcome-based rewards**. Hybrid rollout buffer (Tokens + Latents).
- ▶ **Benefits:** No CoT annotations needed. Scalable and training-efficient.

Key Contributions

1. First RL-Based Approach:

HRPO empowers LLMs to *autonomously* develop latent reasoning capabilities without supervised CoT traces.

2. Novel Gating Design:

Successfully preserves LLMs' generative abilities while progressively integrating continuous representations for internal reasoning.

3. Efficiency & Performance:

- ▶ Eliminates expensive multi-stage training.
- ▶ Outperforms existing baselines on both knowledge-intensive and reasoning-intensive benchmarks.

Methodology

Step 1: Projection (Manifold Alignment)

For input query $x = [x_1, x_2, \dots, x_t]$ and its corresponding token embeddings $E = [e_1, e_2, \dots, e_t]$, the raw hidden states from the LLM output at step t with \hat{h}_t is $\hat{H} = [\hat{h}_1, \hat{h}_2, \dots, \hat{h}_t] = \text{Transformer}(E)$.

Problem: Raw hidden states \hat{h}_t do not match the input embedding space.

Solution: *Weighted Interpolation*. Project \hat{h}_t back using output probabilities p_{t+1} :

$$h_{t+1} = W_e^T \frac{p_{t+1}}{\|p_{t+1}\|}, \text{ with } p_{t+1} = \text{softmax}\left(\frac{\text{Head}(\hat{h}_t)}{\tau}\right),$$

Aligns with native input distribution and **Preserves** differentiability.

Methodology

Step 2: Gating (Hybrid Fusion)

Problem: Need to maintain stochasticity for RL and preserve generation quality.

Solution: *Gating Mechanism*. Mix sampled token \hat{e}_{t+1} with projected state h_{t+1} :

$$\begin{aligned}r_t &= \sigma(W_a \hat{e}_{t+1} + b_a), \\i_t &= \sigma(W_x \hat{e}_{t+1} + b_x), \\a_t &= \exp(-c \cdot \text{softplus}(\Lambda) \odot r_t), \\e_{t+1} &= \begin{cases} a_t \odot \hat{e}_{t+1} + \sqrt{1 - a_t^2} \odot (i_t \odot h_{t+1}) & t \in \text{think}, \\ \hat{e}_{t+1} & t \notin \text{think}, \end{cases}\end{aligned}$$

The gating parameter Λ is selected such that the quantity $a^c = \exp(-c \cdot \text{softplus}(\Lambda))$ is drawn uniformly from $[r_{min}, 0.999]$, with the scalar constant fixed at $c = 8$.

Key Dynamics: Initialization ($a_t \rightarrow 1$), prioritizes discrete tokens initially to protect stability. Gradually learns to mix in latent signals.

Core Logic: Start Discrete (Safe) \rightarrow Evolve to Hybrid (Powerful).

Methodology: HRPO

The Mechanism (GRPO-Style): Maximize expected reward using hybrid outputs (Tokens y + States H), which is $\max_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}, (\hat{y}, H) \sim \pi_{\theta}(\cdot|x)} [r(a, y)]$, the gradient is:

$$\mathbb{E}_{x \sim \mathcal{D}, \{(y_i, H_i)\}_{i=1}^g \sim \pi_{\theta}(\cdot|x)} \left[\frac{1}{g} \sum_{i=1}^g \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \nabla_{\theta} \log \pi_{\theta}(y_{i,t}|x, y_{i,<t}, H_{i,<t}) \hat{A}_{i,t} \right] - \beta \nabla_{\theta} \mathbb{D}_{KL}[\pi_{\theta} \parallel \pi_{\text{ref}}]$$

- **Rollouts:** Sample g hybrid trajectories per input.
- **Reward (r):** Simple outcome-based (1 for correct, 0 otherwise).

Summary: Unifies hybrid reasoning under a simple, efficient RL objective.

Evaluation on Knowledge Benchmarks

Table 1: Evaluation performance of various larger LLMs and trained models on open-domain and multi-hop QA benchmarks. The table reports exact match scores based on top-3 retrieved documents on five datasets: NQ, TriviaQA, HotpotQA, 2WikiMQA and Bamboogle. The upper block reports results for several RAG baselines using the larger Qwen 2.5 7B LLM, while the lower two blocks evaluate smaller Qwen models (1.5B and 3B) trained with different strategies.

	NQ	TriviaQA	HotpotQA	2WikiMQA	Bamboogle	Average
Qwen2.5-7B-Instruct						
QA	0.134	0.408	0.183	0.250	0.120	0.219
CoT	0.048	0.185	0.092	0.111	0.232	0.134
IRCoT	0.224	0.478	0.133	0.149	0.224	0.242
Search-o1	0.151	0.443	0.187	0.176	0.296	0.251
RAG	0.349	0.585	0.299	0.235	0.208	0.335
Qwen2.5-1.5B-Instruct						
SFT	0.094	0.193	0.129	0.210	0.024	0.130
RAG	0.288	0.477	0.228	0.203	0.072	0.254
PPO	0.327	0.527	0.256	0.242	0.184	0.307
GRPO	0.293	0.480	0.202	0.213	0.120	0.261
HRPO (Ours)	0.364	0.553	0.273	0.276	0.216	0.337
Qwen2.5-3B-Instruct						
SFT	0.249	0.292	0.186	0.248	0.112	0.217
RAG	0.348	0.544	0.255	0.226	0.080	0.291
PPO	0.356	0.563	0.304	0.293	0.240	0.351
GRPO	0.381	0.570	0.308	0.303	0.272	0.367
HRPO (Ours)	0.378	0.593	0.316	0.318	0.296	0.380

- **Superior to Other RL Methods:** HRPO consistently outperforms PPO and GRPO across both backbone sizes.
- **Task Strengths:** Largest gains observed on terse queries (NQ) and multi-hop questions (2WikiMQA), proving the efficacy of combining hybrid latent reasoning with retrieval.

Evaluation on STEM Benchmarks

Table 2: Evaluation performance of various larger LLMs and trained models on STEM benchmarks. The table presents accuracy scores on five datasets: GSM8k, MATH, MATH500, MMLU-ST and ARC-C. The upper block reports results for several few-shot baseline LLMs $\geq 7B$, while the lower two blocks evaluate smaller Qwen models (1.5B and 3B) trained with different strategies.

	GSM8k	MATH	MATH500	MMLU-ST	ARC-C	Average
Larger LLMs (Size $\geq 7B$)						
DeepSeekMath-7B	0.642	0.362	0.346	0.565	0.678	0.519
Gemma-2-9B	0.707	0.377	0.364	0.651	0.682	0.556
Qwen2.5-7B	0.854	0.498	0.464	0.723	0.637	0.635
MAmmoTH2-7B	0.684	0.367	0.396	0.624	0.817	0.578
MAmmoTH2-8B	0.704	0.358	0.732	0.642	0.822	0.652
Qwen2.5-1.5B-Instruct						
SFT	0.560	0.300	0.302	0.403	0.602	0.433
Distilled CoT	0.706	0.503	-	-	-	-
PPO	0.694	0.507	0.518	0.566	0.715	0.600
GRPO	0.711	0.502	0.524	0.562	0.737	0.607
HRPO (Ours)	0.720	0.518	0.536	0.569	0.742	0.617
Qwen2.5-3B-Instruct						
SFT	0.670	0.348	0.360	0.454	0.474	0.461
Distilled CoT	0.799	0.575	-	-	-	-
PPO	0.819	0.597	0.604	0.582	0.811	0.682
GRPO	0.834	0.602	0.604	0.601	0.814	0.691
HRPO (Ours)	0.845	0.613	0.630	0.590	0.820	0.700

- **Methodology:** SFT underperforms compared to RL and Distilled CoT, highlighting the value of verifiable rewards in reasoning tasks.
- **Efficiency:** HRPO shows larger gains over GRPO on smaller models (1.5B), proving highly effective for compact backbones.

Different Strategies for Latent Reasoning

Comparison on MATH (Qwen-1.5B)

We compared three methods for integrating hidden states:

1. **Hidden States (Direct):** Degrades generation; results in zero reward.
2. **Interpolation:** Initially similar to HRPO but eventually collapses due to excessive noise.
3. **HRPO (Ours):** Superior training dynamics, fast convergence, and stability comparable to GRPO.

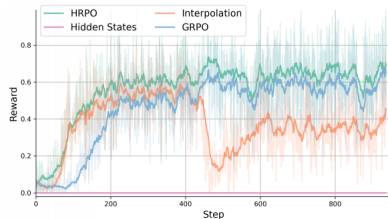


Figure 3: Reward on MATH for Qwen-2.5-1.5B using different latent reasoning strategies.

Latent Ratio

Ratio of Latent Representations

- **Trend:** Hidden ratio increases steadily throughout training.
- **Length:** Completion lengths decrease over time (especially in HRPO).
- **Insight:** Latent representations capture context effectively, enabling shorter, efficient reasoning.

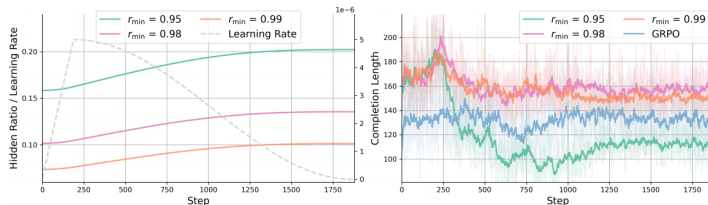


Figure 4: Hidden ratio with varying r_{\min} in $\exp(-c \cdot \text{softplus}(\Lambda))$ and learning rate. We visualize the hidden ratio and completion length for training runs with r_{\min} from $[0.95, 0.98, 0.99]$.

Emergent Patterns

- **Interpretability:** Trajectories remain readable without CoT supervision.
- **Cross-lingual:** Generalizes across linguistic boundaries.
- **Compactness:** More efficient decoding for simple/factual queries.

MMLU Query: This is a true story that happened in Japan. In order to _ the house, a Japanese tore open the walls. Japanese..... What can we learn from the free lizard?

Choices:

- A. It teaches us never to abandon our loved ones. ☒
C. It tells us to take pity on the stuck lizard.

- B. It teaches us to give more help to our loved ones.
D. It encourages us to live even longer.

HRPO (Tokens in Hybrid Rollout)

To solve this problem, we need to understand the meaning of the story and the story it wants to convey. 1. 首先，理解故事的含义：从一只小蜥蜴的角度，它没有移动，而是等待着它的伴侣。直到另一只小蜥蜴出现并带给了它食物，这个小蜥蜴才开始进食。这个小蜥蜴没有放弃它的伴侣，一直在等待。这表明它爱它的伴侣。2. 接下来，我们需要从这个故事中学习什么。故事告诉我们，一个小小的蜥蜴可以爱得像这样，即使它很小，它也可以爱得像一个巨大的生命。这让我们思考，我们如何才能变得更加无私和有爱心。3. 然后，我们需要分析每个选项，看看哪个选项最符合这个故事的意图。A. 它告诉我们要抛弃我们爱的人。这符合故事的意图，因为小蜥蜴没有放弃伴侣，即使它没有移动。B. 它告诉我们给予更多的帮助给我们的爱。这也不完全符合，因为故事中的蜥蜴没有给更多的帮助，它只是等待和照顾它的伴侣。C. 它告诉我们同情被困的小蜥蜴。这不符合故事的意图，因为小蜥蜴没有同情心，它只是等待和照顾它的伴侣。D. 它鼓励我们活得更长。这不符合故事的意图，因为小蜥蜴没有活得更长，它只是等待和照顾它的伴侣。Given this analysis, the correct choice is A. It teaches us never to abandon our loved ones. #### A

Translation for Chinese

1. First, understand the story: From the perspective of a little lizard, it did not move, but waited for its partner, until another little lizard appeared and brought it food, and then the little lizard began to eat. The little lizard did not abandon its partner and kept waiting. This shows that it loves its partner. 2. Next, what do we need to learn from this story. The story tells us that a tiny lizard can love like this, even it is small, it can love like a real living. This makes us think about how we can become more selfless and loving. 3. Then, we need to analyze each option to see which best fits the intention of the story. A. It tells us not to abandon the people we love. This fits the intention of the story as the little lizard did not abandon its partner even if it did not move. B. It tells us to give more help to our love. This is not completely in line, because the lizard in the story did not give more help, it just waited and took care of its partner. C. It tells us to sympathize with the trapped lizard. This does not fit the intention of the story because the little lizard has no sympathy, it just waited and took care of its partner. D. It encourages us to live longer. This does not fit the intention of the story, as the little lizard did not live longer, it was just waiting and caring for its mate.

Figure 6: Example cross-lingual reasoning (English-Chinese) and its translation for HRPO.

Conclusion

- ▶ Proposed HRPO (Hybrid Reasoning Policy Optimization).
- ▶ **Novel Framework:** Unifies **discrete token sampling** with **continuous latent representations** via a **learnable gating mechanism**.
- ▶ **Mechanism:** Incentivizes LLMs to refine reasoning strategies hybridly by gradually incorporating hidden features.
- ▶ **Performance:** Consistently outperforms both SFT and standard RL baselines across diverse scenarios.
- ▶ **Future Work:** Ensures stable training for hybrid latent reasoning.

Thank you!