NeurIPS 2025

Hybrid Latent Reasoning via Reinforcement Learning

Zhenrui Yue[1], **Bowen Jin**[1], **Huimin Zeng**[1], **Honglei Zhuang**[2], **Zhen Qin**[2], **Jinsung Yoon**[2], **Lanyu Shang**[3], **Jiawei Han**[1], **Dong Wang**[1]
[1]University of Illinois Urbana-Champaign, [2]Google, [3]LMU
{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 $x = [x_1, x_2, ..., x_t]$ and its corresponding token embeddings $E = [e_1, e_2, ..., e_t]$, the raw hidden states from the LLM output at step t with $\hat{h}_t$ is $\hat{H} = [\hat{h}_1, \hat{h}_2, ..., \hat{h}_t] = \text{Transformer}(E)$.

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}: $h_{t+1} = W_e^T \frac{p_{t+1}}{\|p_{t+1}\|}$, 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}:

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

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

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

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

Methodology: HRPO

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

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

- **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.

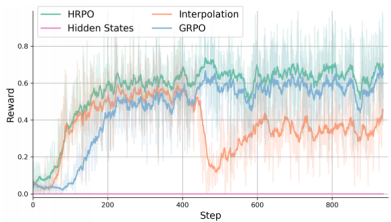

Figure 3: Reward on MATH for Qwen-2.5-1.5B using different latent reasoning strategies.

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.

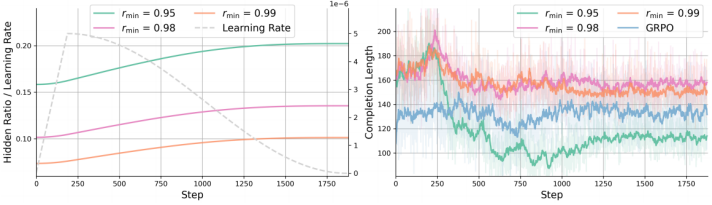

Figure 4: Hidden ratio with varying $r_{\min}$ in $\exp(-c \cdot \texttt{softplus}(\Lambda))$ and learning rate. We visualize the hidden ratio and completion length for training runs with $r_{\min}$ from $[0.95, 0.98, 0.99]$.

Emergent Patterns

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

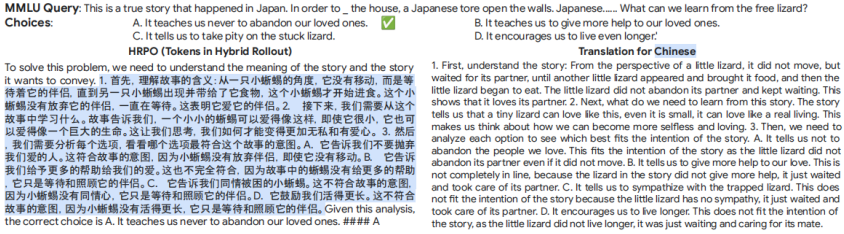

Figure 6: Example cross-lingual reasoning (English-Chinese) and its translation for HRPO.

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!