ATPO: Adaptive Tree Policy Optimization For Multi-turn Medical Dialogue

Ruike Cao[1,2,*] **Shaojie Bai[1,3,*]** Fugen Yao[1,†] Liang Dong[1] Jian Xu[1] Li Xiao[2]

[1]Qwen Applications Business Group, Alibaba Group

[2]University of Science and Technology of China

[3]Zhejiang University

***Equal contribution**

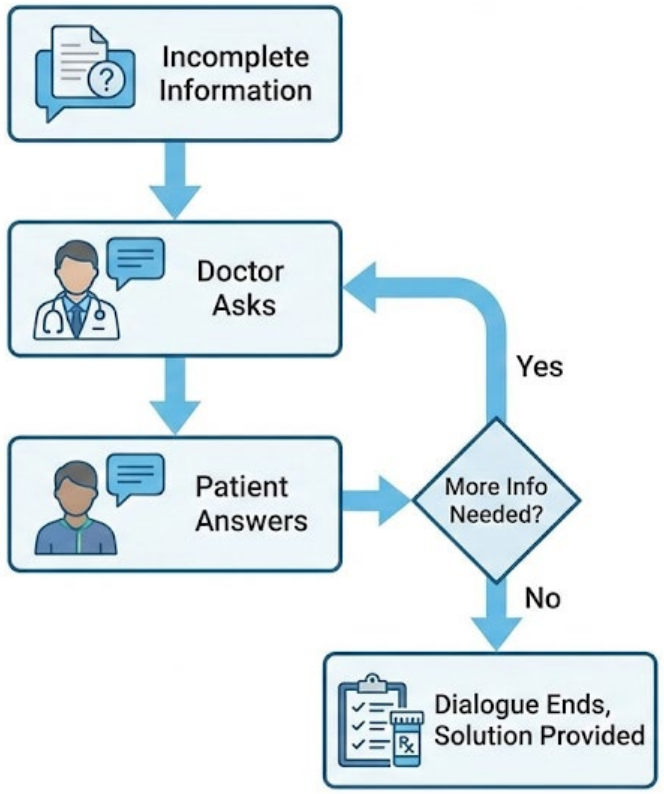
[†]Corresponding author: fugen.yfg@alibaba-inc.com

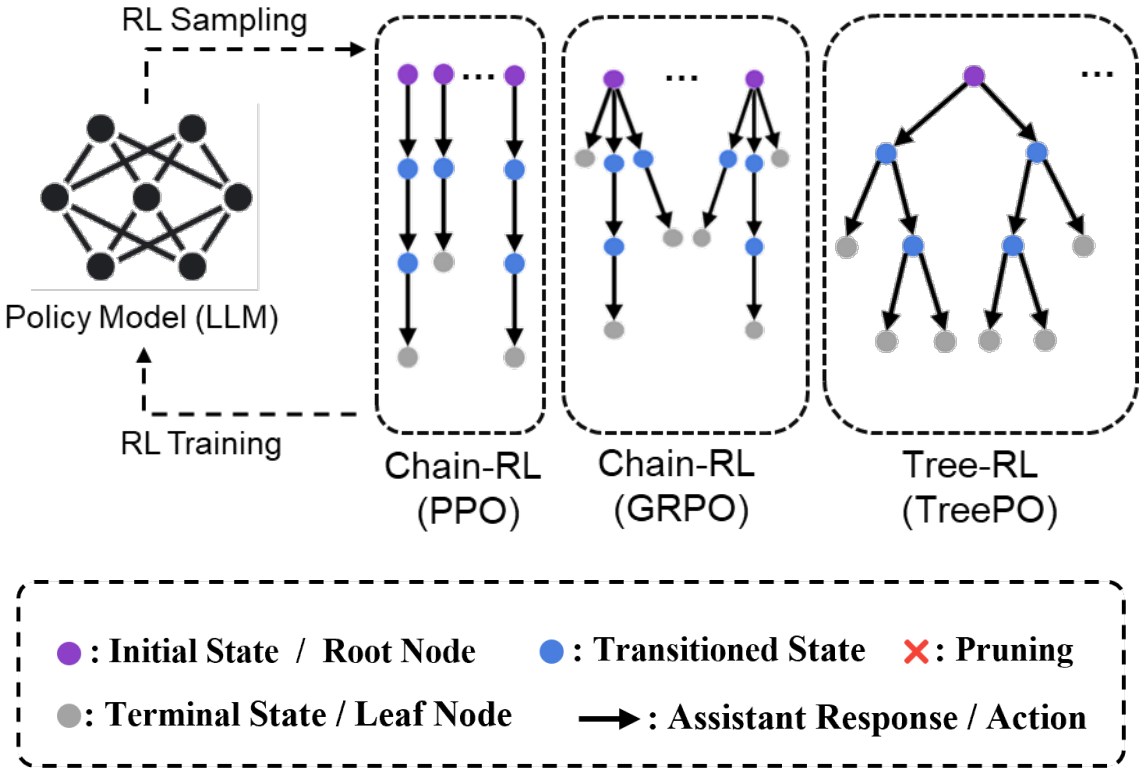
Code: https://github.com/Quark-Medical/ATPO

Reporter: Shaojie Bai

2/28/2026

Background & Motivation



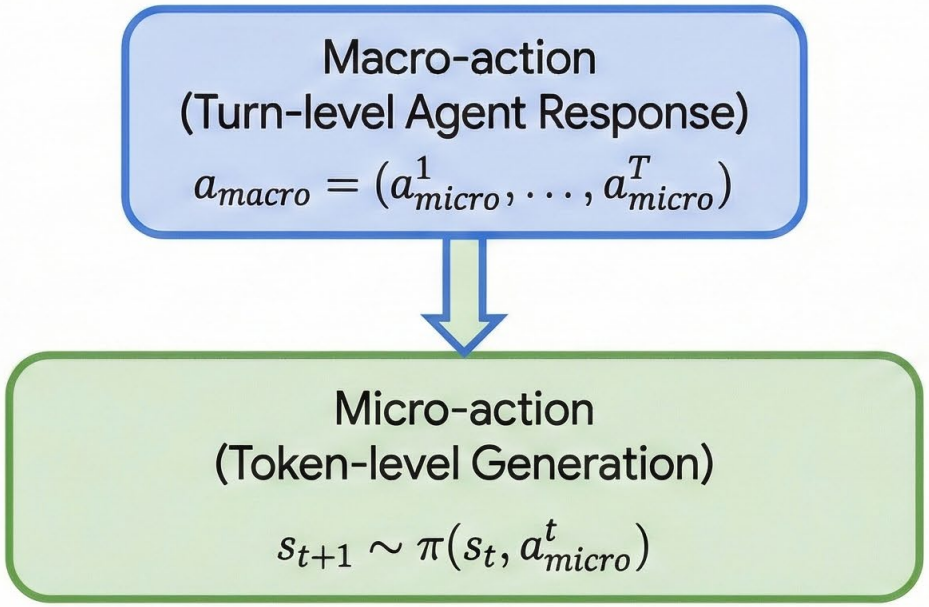


- **Multi-turn medical dialogue:** lack of inquiry capabilities.

- **PPO:** unstable value estimation
- **GRPO:** long-horizon credit assignment
- **TreePO:** low computational efficiency

Formulation & Tree Search Framework

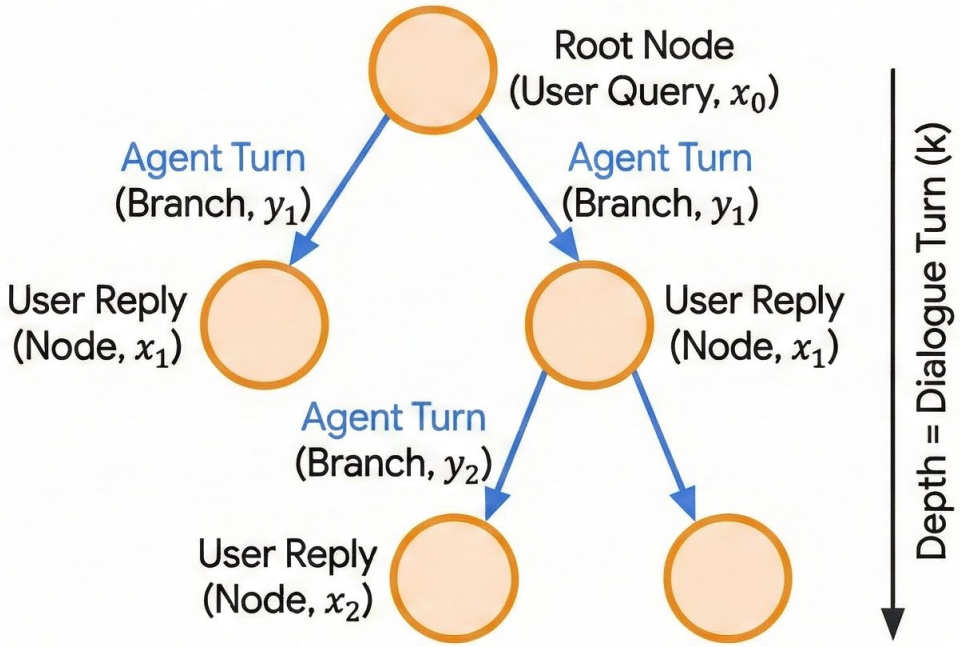
Hierarchical MDP (H-MDP)



- Formulates dialogue as a hierarchy of turn-level and token-level decisions.

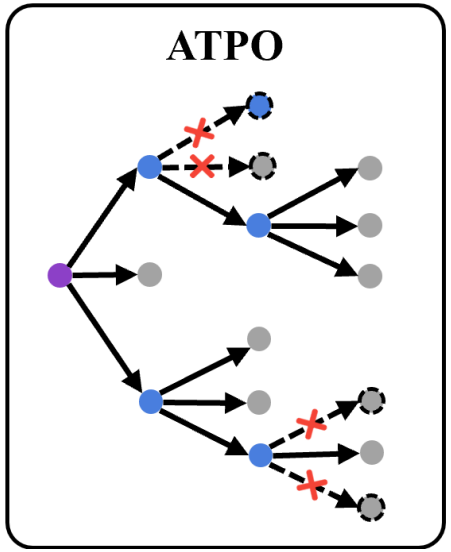
Transition: $\mathcal{P}(s'|s, a_{macro}) = \prod_{t=1}^{T} \mathcal{P}(s_{t+1}|s_t, a^t_{micro})$

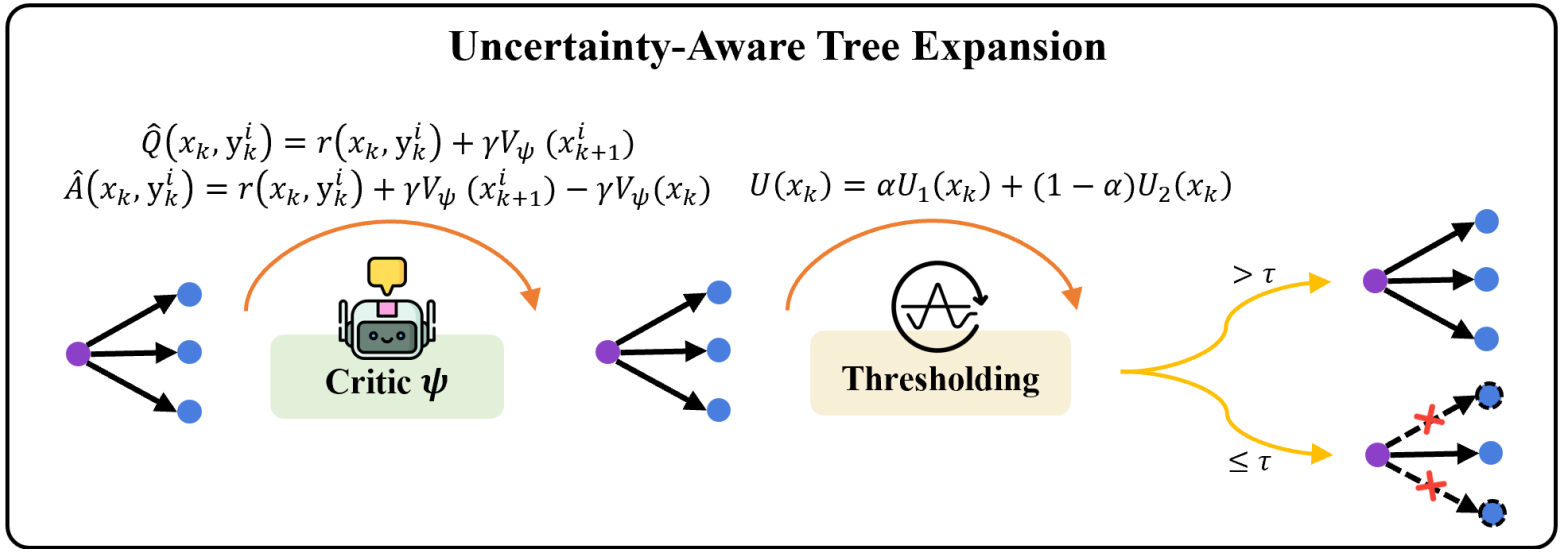
Tree-Based Exploration (ATPO)



- Branch: Agent's message (action)
- Node: User's message (state)
- Depth: Number of interaction turns

Proposed Method - ATPO





● : Initial State / Root Node ● : Transitioned State ✕ : Pruning

● : Terminal State / Leaf Node → : Assistant Response / Action

State Transition

x_0 ⇨ → y_0 ⇨ User Response and $r(x_0, y_0)$ ⇨ x_1

Uncertainty $U(x_k)$

$$U(x_k) = \alpha U_1(x_k) + (1-\alpha) U_x(x_k)$$

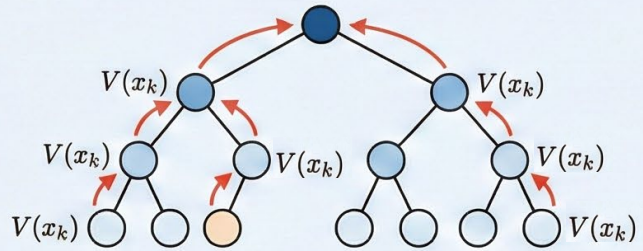
$U_1(x_k)$: Bellman Error
Critic Training Value

$U_2(x_k)$: Action-Value Variance
Sampling Diversity

Proposed Method - ATPO

Model Updating: Two-Step Process

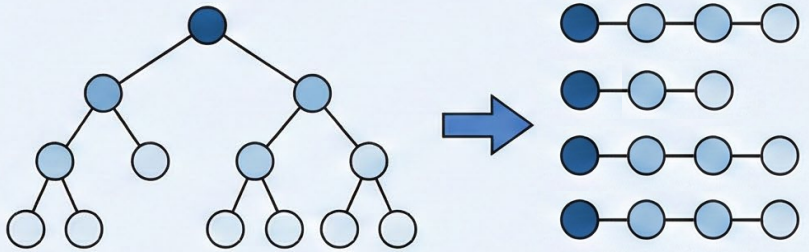
1. Value Traceback (GAE on Tree)



Backward pass to compute state values and advantages.

$$A^{\text{GAE}}(x_k) = \sum \gamma^l \delta_{k+l}$$

2. Tree Decomposition

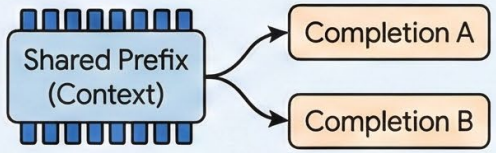


Update Policy via PPC loss utilized on these paths:

$$\mathcal{L}^{\text{CLIP}}(\theta) = \mathbb{E}_{\tau \sim \text{Tree}} \left[\sum_t \min(r_t A_t, \text{clip}(\dots) A_t) \right]$$

System Efficiency: Three Key Engineering Optimizations

KV Cache Reuse

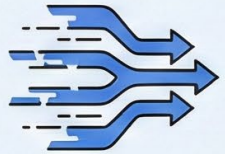


Avoids recomputing shared history for distinct branches.

Asynchronous Search

Parallelizes generation and evaluation to maximize utilization.

High Throughput (vLLM)



Enables massive-scale tree exploration in feasible time.

Experimental Setup

- **Environment & Task:** A multi-turn clinical case reasoning environment. An Assistant Agent interactively queries a User Simulator (Qwen3-8B) to obtain missing facts and resolve clinical questions.
- **Base Models:** Evaluated on the Qwen3 series across three different sizes (Qwen3-1.7B, Qwen3-4B, and Qwen3-8B).
- **Baselines & Comparisons:**
  - Prompting & Fine-Tuning: Zero-shot Prompting (Direct & MEDIQ) , Supervised Fine-Tuning (SFT), and Dynamic Fine-Tuning (DFT).
  - RL Baselines: Critic-based methods (PPO MDP & H-MDP) and critic-free methods (GRPO & TreePO).
  - State-of-the-Art LLMs: Compared against strong expert models like GPT-4o and Gemini-2.5-Pro.

Key Experimental Results (Accuracy)

Table 1: Performance comparison (%) on **MedicalExam**, **MedQA**, and **MedMCQA**. **Bold** indicates the best performance, underlined the second-best (excluding GPT-4o and Gemini-2.5-Pro).

Model	Method Type	Method Name	MedicalExam	MedQA	MedMCQA
Qwen3-1.7B	Prompt	Direct	35.07 ± 1.12	34.05 ± 0.38	32.54 ± 0.49
		MEDIQ	34.00 ± 2.26	34.20 ± 0.75	32.35 ± 1.73
	SFT	DFT	29.07 ± 1.46	28.38 ± 0.80	21.08 ± 0.90
		SFT	32.27 ± 4.77	33.42 ± 0.95	28.10 ± 2.32
	SFT+RL	PPO (MDP)	39.33 ± 4.01	38.64 ± 1.17	35.37 ± 0.80
		PPO (H-MDP)	39.33 ± 2.79	39.08 ± 1.85	34.89 ± 1.00
		GRPO	42.93 ± 1.80	41.17 ± 0.64	36.57 ± 3.26
		TreePO	43.33 ± 1.56	42.05 ± 1.03	38.47 ± 2.00
		ATPO (U_1)	**45.73 ± 1.53**	42.54 ± 0.39	38.66 ± 0.66
		ATPO ($U_1 + U_2$)	43.20 ± 1.85	**42.87 ± 0.77**	**39.93 ± 1.05**
Qwen3-4B	Prompt	Direct	48.13 ± 0.87	44.94 ± 0.35	41.53 ± 0.39
		MEDIQ	45.87 ± 1.20	40.11 ± 0.60	31.64 ± 1.41
	SFT	DFT	43.07 ± 1.61	41.72 ± 1.27	33.28 ± 1.68
		SFT	48.93 ± 2.14	47.15 ± 1.01	39.18 ± 1.22
	SFT+RL	PPO (MDP)	50.13 ± 2.80	50.60 ± 0.90	42.50 ± 0.84
		PPO (H-MDP)	52.40 ± 2.24	48.58 ± 1.48	43.32 ± 2.22
		GRPO	53.87 ± 2.08	51.17 ± 1.08	43.84 ± 0.78
		TreePO	56.13 ± 0.99	53.74 ± 0.56	45.22 ± 0.65
		ATPO (U_1)	56.80 ± 1.28	53.15 ± 0.55	**46.23 ± 1.25**
		ATPO ($U_1 + U_2$)	**59.73 ± 2.61**	**55.47 ± 0.99**	45.93 ± 1.13
Qwen3-8B	Prompt	Direct	52.40 ± 0.37	45.22 ± 0.34	46.16 ± 1.04
		MEDIQ	51.87 ± 3.69	46.03 ± 0.75	41.60 ± 0.91
	SFT	DFT	51.86 ± 3.63	48.80 ± 1.30	42.20 ± 0.83
		SFT	55.87 ± 0.30	53.75 ± 1.18	46.87 ± 1.74
	SFT+RL	PPO (MDP)	59.20 ± 3.84	57.38 ± 0.84	50.00 ± 0.81
		PPO (H-MDP)	59.07 ± 3.15	57.81 ± 1.29	51.98 ± 0.67
		GRPO	60.93 ± 1.86	57.92 ± 0.68	51.12 ± 1.29
		TreePO	65.33 ± 3.09	61.81 ± 0.90	**54.74 ± 1.99**
		ATPO (U_1)	65.52 ± 3.12	62.57 ± 0.41	53.22 ± 1.30
		ATPO ($U_1 + U_2$)	**65.87 ± 3.72**	**64.07 ± 0.43**	53.66 ± 1.52
GPT-4o	Prompt	MEDIQ	64.00 ± 3.53	63.15 ± 0.82	53.03 ± 0.89
Gemini-2.5-Pro	Prompt	MEDIQ	74.33 ± 2.53	68.69 ± 0.61	63.31 ± 1.37

- **SOTA Accuracy:** ATPO consistently outperforms strong RL baselines (PPO, GRPO, TreePO) across all datasets.
- **Surpassing GPT-4o:** ATPO on Qwen3-8B beats the much larger GPT-4o (+0.92% on MedQA).
- **Metric Synergy:** Combining Bellman error (U_1) and value variance (U_2) yields the absolute best performance.

Key Experimental Results (Efficiency)

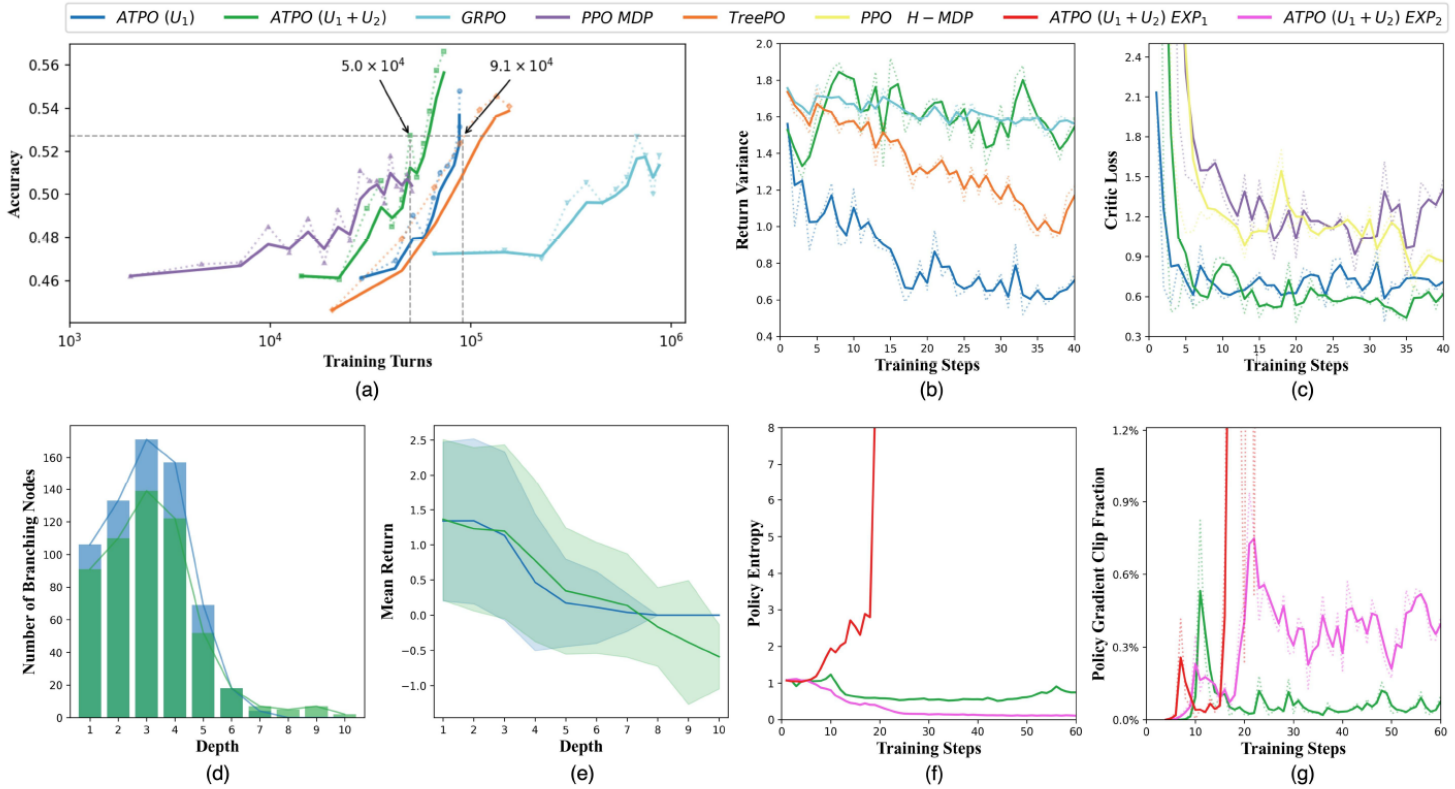


Figure 2: Analysis of the ATPO algorithm on Qwen3-4B. (a) Training efficiency and performance comparison of various algorithms, plotting accuracy against the number of generated turns. (b), (c) Return variance and critic loss for ATPO and baseline methods. (d), (e) Distribution of branching nodes and returns by depth for samples from ATPO at a representative training step. (f), (g) Stability analysis of ATPO with and without visit-count-based down-weighting.

- **High Sample Efficiency:** ATPO reaches 52.7% accuracy using only 55% of the training turns required by TreePO.
- **Enhanced Critic Training:** Dual metrics ($U_1 + U_2$) significantly lower critic loss while maintaining diverse exploration.
- **Deeper & Balanced Search:** Uncertainty-based adaptive pruning prevents exponential early-node growth, enabling much deeper exploration.

Conclusion & Future Work

- **Conclusions:**
  - Proposed **ATPO**, an adaptive tree search method guiding exploration via state-uncertainty evaluation.
  - Achieves superior performance and higher sample efficiency, outperforming strong baselines (TreePO, GRPO) and GPT-4o.
  - Broadly applicable beyond medical domain, such as open-ended dialogues and tool use.
- **Future Work**
  - Developing a **learnable, soft control policy** to dynamically adapt expansion thresholds.
  - **Refining credit assignment** within the H-MDP framework beyond uniform token-level cloning.

Q&A

Thank You!

Code & Datasets available at:
https://github.com/Quark-Medical/ATPO