IAPO: Information-Aware Policy Optimization for Token-Efficient Reasoning

Yinhan He[1], Yaochen Zhu[1], Mingjia Shi[1], Wendy Zheng[1], Lin Su[2], Xiaoqing Wang[2], Qi Guo[2], Jundong Li[1]
[1]University of Virginia, [2]LinkedIn Inc.

Motivation and Challenges

Motivation: With RL-based post-training (e.g., GRPO), LLMs achieve strong reasoning accuracy, yet produce excessively verbose reasoning chains. DeepSeekR1-Distilled-Qwen-1.5B generates 1,658 tokens on average vs. 264 tokens by a human on MATH-500 problems—both achieving perfect accuracy. Such verbosity inflates inference latency and cost, which scale quadratically with sequence length.

Challenges:

- ***Content-Agnostic Advantage Shaping***: Existing token-efficient methods (length-based and position-based) assign advantages by completion length or token position, ignoring whether a token actually contributes to the correct answer.
- ***Computational Intractability***: Quantifying a token's contribution (conditional mutual information (MI)) to the final answer requires two forward passes per token, yielding prohibitive complexity for long reasoning sequences.

Contributions

- ***Information-Theoretic Metric:*** First framework to assign token-wise advantages based on each token's conditional mutual information (MI) with the final answer, providing a principled mechanism for identifying informative vs. redundant tokens.
- ***Efficient MI Estimation:*** An early-exit-based conditional MI estimator with KV-cache preloading and chunk-wise forwarding, reducing complexity from cubic to quadratic in the sequence length.
- ***Strong Empirical Results:*** Consistent improvements across 3 LLM scales (0.5B, 1.5B, 7B) and 3 math reasoning datasets, achieving up to 47% token reduction while preserving or improving accuracy (e.g., 1.16×10^{-2} Ratio@16 on GSM8K **with 7B model).**

Methodology

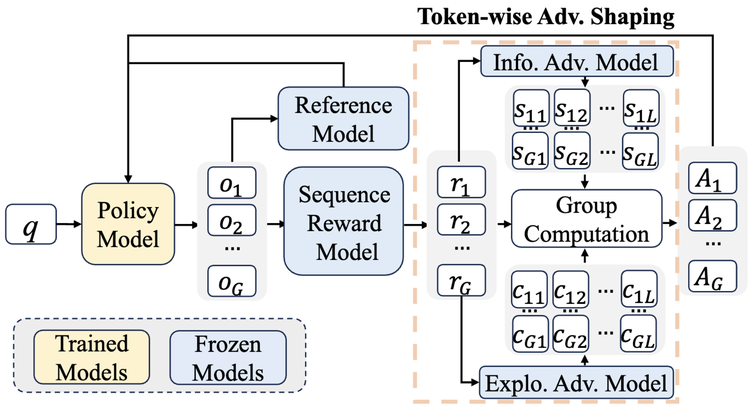

Figure 1: An overview of the proposed IAPO framework.

- ***Information-Aware Advantage Shaping:***
  - ***Informativeness Level:*** Each token's contribution is measured via conditional MI with the final answer: $s_{i,t} = I(y_i; o_t | q_i, o_{<t})$, quantifying uncertainty reduction about the answer after observing that token.
  - ***Exploration Adjustment***: Token-level exploration term $c_{i,t} = \pm\pi_{\theta_{old}}(o_{i,t}|q, o_{i,<t})$: positive for correct completions (suppress exploration), negative for incorrect (encourage exploration).
  - ***Token-Wise Advantage***: $\tilde{A}_{i,t} = \mathrm{norm}(r_i, r) + \alpha \cdot \mathrm{norm}(s_{i,t}, s_i) + \beta \cdot \mathrm{norm}(c_{i,t}, c_i)$, combining sequence-level reward, token-level informativeness, and token-level exploration.
- ***Efficient Conditional MI Estimation:***
  - ***Early-Exit Estimator:*** Approximates $I(y; o_t|q, o_{<t}) = H(y|q, o_{<t}) - H(y|q, o_{\le t})$ by appending a postfix prompt (e.g., "</think><answer>") to elicit answer distributions at each prefix.
  - ***KV-Cache Preloading:*** Single forward pass caches all prefix states; entropy evaluations reuse cached KV states with only the short postfix, significantly reducing complexity per token.
  - ***Chunk-wise Forwarding:*** Batches MI estimation across contiguous token chunks, amortizing costs by a factor.

Experiments

Method	P@16	P@32	L@16	L@32	R@16	R@32
Dataset GSM8K (Cobbe et al., 2021)						
Qwen2.5-0.5B-Instruct	0.4261±0.0099	0.5722±0.0081	148.04±3.51	149.64±6.01	2.88×10^{-3}	3.82×10^{-3}
+ DAPO (Yu et al., 2025)	0.8605±0.0006	0.9050±0.0004	172.95±0.27	172.53±0.09	4.98×10^{-3}	5.25×10^{-3}
+ GFPO (Shrivastava et al., 2025)	0.8590±0.0056	0.9052±0.0043	216.26±0.31	215.99±0.12	3.97×10^{-3}	4.19×10^{-3}
+ GTPO (Tan et al., 2025)	0.8519±0.0041	0.9085±0.0028	243.17±0.26	243.21±0.32	3.50×10^{-3}	3.74×10^{-3}
+ S-GRPO (Lee & Tong, 2025)	0.8489±0.0077	0.9017±0.0009	159.70±0.13	159.49±0.20	5.32×10^{-3}	5.65×10^{-3}
+ IAPO (ours)	0.8519±0.0084	0.8979±0.0068	150.37±0.22	150.15±0.24	$\mathbf{5.67\times10^{-3}}$	$\mathbf{5.98\times10^{-3}}$
Qwen2.5-1.5B-Instruct	0.8160±0.0090	0.8941±0.0029	152.66±1.53	152.59±3.76	5.35×10^{-3}	5.86×10^{-3}
+ DAPO (Yu et al., 2025)	0.9479±0.0059	0.9664±0.0050	169.31±0.37	169.80±0.26	5.60×10^{-3}	5.69×10^{-3}
+ GFPO (Shrivastava et al., 2025)	0.9510±0.0029	0.9725±0.0022	203.85±0.65	204.05±0.58	4.67×10^{-3}	4.77×10^{-3}
+ GTPO (Tan et al., 2025)	0.9497±0.0028	0.9674±0.0012	262.63±0.62	262.49±0.78	3.62×10^{-3}	3.69×10^{-3}
+ S-GRPO (Lee & Tong, 2025)	0.9558±0.0032	0.9735±0.0027	181.07±0.08	180.85±0.16	5.28×10^{-3}	5.38×10^{-3}
+ IAPO (ours)	0.9512±0.0034	0.9707±0.0028	163.32±0.46	163.51±0.56	$\mathbf{5.82\times10^{-3}}$	$\mathbf{5.94\times10^{-3}}$
Qwen2.5-7B-Instruct	0.9793±0.0009	0.9851±0.0016	157.08±1.76	156.43±1.05	6.23×10^{-3}	6.30×10^{-3}
+ DAPO (Yu et al., 2025)	0.9778±0.0026	0.9816±0.0026	160.36±0.73	159.83±0.21	6.10×10^{-3}	6.14×10^{-3}
+ GFPO (Shrivastava et al., 2025)	0.9798±0.0016	0.9853±0.0009	160.36±0.99	160.50±0.57	6.11×10^{-3}	6.14×10^{-3}
+ GTPO (Tan et al., 2025)	0.9765±0.0006	0.9826±0.0006	192.99±0.10	192.85±0.10	5.06×10^{-3}	5.10×10^{-3}
+ S-GRPO (Lee & Tong, 2025)	0.9790±0.0020	0.9843±0.0004	147.67±0.70	147.40±0.42	6.63×10^{-3}	6.68×10^{-3}
+ IAPO (ours)	0.9735±0.0012	0.9805±0.0009	84.03±0.12	83.97±0.04	$\mathbf{1.16\times10^{-2}}$	$\mathbf{1.17\times10^{-2}}$

IAPO achieves the best token-efficiency (Ratio@k) across nearly all settings while maintaining competitive or superior accuracy. On GSM8K with Qwen2.5-7B, IAPO reduces reasoning length by 47% compared to DAPO with no accuracy loss. **More results from paper**: Wall-clock inference time is reduced by up to 17.7%. Ablation studies confirm that both the informativeness level term (conditional MI) and the early-exit estimator contribute to IAPO's token efficiency gains.

Acknowledgements

This work was supported in part by the National Science Foundation (NSF) under Grants IIS-2144209, IIS-2223769, BCS-2228534, and CMMI-2411248, and by the Office of Naval Research (ONR) under Grant N000142412636.