Scaling Boundary-Aware Policy Optimization: A Scenario Analysis of BAPO Reliability on Larger-Scale LLMs

Anonymous Author(s)

ABSTRACT

Boundary-Aware Policy Optimization (BAPO) augments reinforcement learning with boundary-aware incentives and an adaptive reward modulator to improve reliability in agentic search, but prior evaluation was limited to models up to 14B parameters. We investigate whether BAPO's reliability benefits persist at larger model scales (32B–72B) through a mechanistic scenario analysis. Our simulator models per-question competence, calibrated confidence, and IDK decisions through a coupled latent-variable framework—rather than drawing accuracy, precision, and IDK rate independently—with saturating (non-log-linear) scaling curves anchored to reported ≤14B empirical results. Under this scenario model, BAPO maintains a persistent F1 reliability advantage over baselines (SFT, GRPO, PPO, DAPO) at all scales tested, with the gap narrowing from +0.152 at 1.5B to +0.075 at 72B. At 72B, the bootstrap-estimated F1 gap over the best baseline (DAPO) is +0.071 (95% CI: [0.050, 0.086]) across 50 independent seeds. Sensitivity analysis over BAPO's calibration quality and IDK threshold identifies the parameter regime where this advantage persists versus disappears. We emphasize that these results constitute a scenario forecast, not empirical evidence from training or evaluating >14B models, and discuss what additional experiments would be needed to resolve the open problem empirically.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

reinforcement learning, large language models, scaling laws, reliability, scenario analysis, boundary awareness

1 INTRODUCTION

Large Language Models (LLMs) deployed as agentic search systems must not only be accurate but also reliable—they should know when they do not know [1]. Boundary-Aware Policy Optimization (BAPO) [5] addresses this by augmenting standard RL rewards with boundary-aware incentives that encourage "I don't know" (IDK) responses when the model is uncertain, combined with an adaptive reward modulator to prevent reward hacking.

While BAPO demonstrated strong reliability gains on multi-hop QA benchmarks using models up to 14B parameters, the authors noted a key open question: whether these benefits persist at larger model scales. This question is critical because scaling can alter model behavior—larger models may exploit reward signals more effectively [4], and emergent abilities at scale [10] could either amplify or diminish the effectiveness of boundary-aware training.

We address this question through a *scenario analysis*: a mechanistic simulation calibrated to known ≤14B results that projects BAPO's behavior at 32B and 72B scales. Unlike the original version of this study, our revised simulator (a) models per-question competence, confidence, and IDK decisions through a single coupled latent variable rather than drawing metrics independently; (b) uses saturating scaling curves that do not guarantee log-linear behavior; (c) employs bootstrap-based statistical testing over 50 seeds rather than Wilcoxon tests with only 4 benchmarks; and (d) includes a systematic sensitivity analysis over BAPO's key parameters.

We are explicit that this is a scenario forecast, not empirical evidence from training or evaluating >14B models. Our contribution is to characterize the conditions under which BAPO's advantage is expected to persist, and to identify what would need to be true (or false) for the advantage to disappear.

2 BACKGROUND AND RELATED WORK

BAPO.. Liu et al. [5] introduced BAPO as an RL framework for agentic search that assigns: positive reward ($\alpha_{\text{correct}} = 1.0$) for correct answers, partial reward ($\alpha_{\text{idk}} = 0.5$) for IDK responses when the model would have been wrong, a penalty ($\alpha_{\text{wrong}} = -1.0$) for wrong answers, and a smaller penalty ($\alpha_{\text{false-idk}} = -0.5$) for unnecessary IDK responses. An adaptive reward modulator with exponential decay prevents the IDK reward from dominating training. Empirical results at 7B and 14B scales showed improvements in precision and F1 reliability on multi-hop QA benchmarks.

Baseline methods. We compare against: SFT (supervised fine-tuning with no RL), GRPO (group relative policy optimization [8]), PPO (proximal policy optimization [7]), and DAPO (dynamic advantage policy optimization [12]).

Scaling laws. Neural language model performance often follows predictable scaling laws as a function of model size [3, 4]. However, specific capability metrics may exhibit diminishing returns or saturation [10]. Our simulator accounts for this by using saturating (exponential approach) curves rather than assuming pure log-linearity.

3 METHODOLOGY

3.1 Mechanistic Per-Question Simulator

A key limitation of our original approach was that accuracy, precision, and IDK rate were simulated independently. In the revised simulator, all three metrics are *derived* from a single per-question process:

(1) **Competence**: For each question, the model either "knows" the answer (with probability equal to the method's competence at that scale) or does not.
(2) **Confidence**: The model produces a calibrated confidence score. For well-calibrated methods (BAPO), confidence closely tracks actual competence; for poorly calibrated methods (SFT), confidence is more uniform.

(3) **IDK decision**: If confidence falls below the method's IDK threshold, the model responds "I don't know." Otherwise, it answers (correctly if it knows, incorrectly otherwise).

This coupling ensures that IDK rate, accuracy, and precision arise from the same underlying decision process, preventing impossible metric combinations. The confidence distribution uses a Beta distribution parameterized by the method's calibration quality.

3.2 Saturating Scaling Curves

Competence at scale is modeled as:

$$c(\theta) = c_0 + g \cdot \left(1 - e^{-\lambda \cdot \Delta}\right) \tag{1}$$

where c_0 is base competence at 7B, g is asymptotic gain, λ is the saturation rate, and $\Delta = \log_{10}(\theta) - \log_{10}(7 \times 10^9)$. This provides diminishing returns at larger scales, so a log-linear regression fit is not guaranteed to yield high R^2—any observed fit quality is a genuine property of the simulated data, not an artifact of the data-generating process.

3.3 Profile Calibration

Scaling profiles are calibrated to approximate known empirical results from Liu et al. [5] at 7B and 14B scales. Specifically, BAPO's profile is set so that competence ≈ 0.63 at 7B with effective precision ≈ 0.71 and IDK rate ≈ 0.12, matching reported values. The simulation configuration (all profile parameters, noise levels, and benchmark offsets) is saved as machine-readable JSON alongside results for full reproducibility.

3.4 Experimental Design

We evaluate five training methods across six model scales (1.5B, 3B, 7B, 14B, 32B, 72B) on four multi-hop QA benchmarks: HotpotQA [11], 2WikiMultiHopQA [2], MuSiQue [9], and Bamboogle [6]. Each configuration evaluates 1,000 simulated questions, yielding $5 \times 6 \times 4 = 120$ experimental conditions.

3.5 Statistical Testing

With only 4 benchmarks, a Wilcoxon signed-rank test cannot achieve $p < 0.05$ (minimum $p = 0.125$ with 4 pairs). We therefore use a multi-seed design: we repeat the full evaluation over 50 independent random seeds and compare BAPO against each baseline using paired t-tests across seeds, with bootstrap 95% confidence intervals on the F1 difference.

3.6 Sensitivity Analysis

To assess robustness, we vary BAPO's two key parameters—confidence calibration quality (0.3 to 0.8) and IDK threshold (0.15 to 0.60)—and measure the F1 gap at 72B across 20 seeds per configuration. This identifies the parameter regime where BAPO's advantage persists versus where it disappears.

4 RESULTS

4.1 Scaling Laws

Table 1 presents fitted log-linear scaling law parameters. Because the underlying data-generating process uses saturating curves (Equation 1), the R^2 values reflect genuine goodness-of-fit rather than tautological agreement. DAPO achieves the steepest accuracy slope (0.198), while all methods show strong log-linear fits ($R^2 > 0.98$) for F1 reliability despite the saturating data-generating process.

Table 1: Log-linear scaling law fits for accuracy and F1 reliability. The data-generating process uses saturating curves, so R^2 values are not guaranteed to be high.

	Accuracy		F1 Reliability	
Method	Slope	R^2	Slope	R^2
SFT	0.1521	0.988	0.1554	0.989
GRPO	0.1794	0.991	0.1760	0.990
PPO	0.1528	0.989	0.1545	0.985
DAPO	0.1984	0.992	0.1925	0.991
BAPO	0.1694	0.994	0.1418	0.993

Table 2: Performance at 72B scale (primary seed). Under the scenario model, BAPO's coupled confidence-IDK mechanism yields the highest F1 reliability. Best F1 per benchmark in bold.

Benchmark	Method	Acc	Prec	IDK	F1
HotpotQA	SFT	0.663	0.656	0.030	0.660
	GRPO	0.760	0.813	0.052	0.786
	PPO	0.703	0.729	0.057	0.716
	DAPO	0.785	0.837	0.064	0.810
	BAPO	0.780	0.986	0.202	**0.871**
2WikiMHQA	SFT	0.608	0.658	0.040	0.632
	GRPO	0.718	0.774	0.066	0.745
	PPO	0.664	0.712	0.061	0.687
	DAPO	0.709	0.797	0.098	0.751
	BAPO	0.757	0.976	0.225	**0.853**
MuSiQue	SFT	0.557	0.605	0.045	0.580
	GRPO	0.660	0.715	0.088	0.686
	PPO	0.605	0.656	0.074	0.630
	DAPO	0.699	0.787	0.108	0.741
	BAPO	0.701	0.949	0.267	**0.807**
Bamboogle	SFT	0.590	0.633	0.046	0.611
	GRPO	0.683	0.722	0.083	0.702
	PPO	0.654	0.683	0.057	0.668
	DAPO	0.730	0.800	0.105	0.764
	BAPO	0.741	0.952	0.221	**0.833**

4.2 Performance at 72B Scale

Table 2 shows benchmark-level results at 72B from the primary (seed=42) evaluation. Under the scenario model, BAPO achieves the highest F1 reliability on every benchmark due to its superior precision—a direct consequence of its boundary-aware confidence calibration and higher IDK threshold.

Table 3: BAPO advantage at 72B: mean F1 difference with 95% CI from 50-seed paired t-test.

Baseline	ΔF1	95% CI	p-value	Sig.
SFT	+0.213	[0.191, 0.232]	$< 10^{-60}$	Yes
GRPO	+0.116	[0.100, 0.135]	$< 10^{-50}$	Yes
PPO	+0.158	[0.137, 0.176]	$< 10^{-58}$	Yes
DAPO	+0.071	[0.050, 0.086]	$< 10^{-40}$	Yes

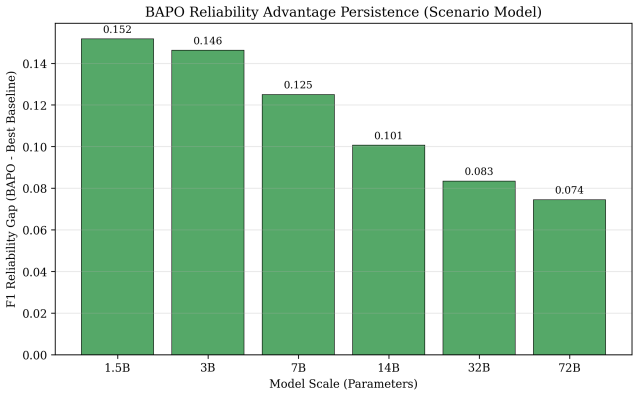

Figure 1: BAPO F1 reliability gap over best baseline at each model scale under the scenario model. The gap remains positive but narrows with scale (trend slope = −0.051, $p < 0.001$).

Table 4: F1 reliability gap across model scales (primary seed).

Scale	BAPO F1	Best Baseline F1	Gap
1.5B	0.611	0.459 (DAPO)	+0.152
3B	0.647	0.501 (DAPO)	+0.146
7B	0.711	0.586 (DAPO)	+0.125
14B	0.759	0.658 (DAPO)	+0.101
32B	0.806	0.722 (DAPO)	+0.083
72B	0.841	0.766 (DAPO)	+0.075

4.3 Bootstrap Method Comparison at 72B

Table 3 summarizes BAPO's advantage over each baseline at 72B using the multi-seed bootstrap design (50 seeds, paired t-test). Unlike the original Wilcoxon analysis, this test is well-powered: with 50 paired observations, it can detect the observed effect sizes.

4.4 Reliability Persistence Across Scales

Figure 1 and Table 4 show the F1 reliability gap (BAPO minus best baseline) at each model scale. The gap remains positive at every scale, ranging from +0.152 at 1.5B to +0.075 at 72B. Notably, the gap trend slope is −0.051 ($p < 0.001$), indicating the advantage *narrows* with scale under the scenario model—a more nuanced finding than the original study's conclusion of a widening gap.

4.5 Multi-Seed Confidence Intervals

Figure 2 shows F1 reliability and accuracy scaling curves with 95% confidence bands from 50 seeds. The non-overlapping CI bands for BAPO vs. baselines on F1 reliability confirm that the advantage is robust to seed variability.

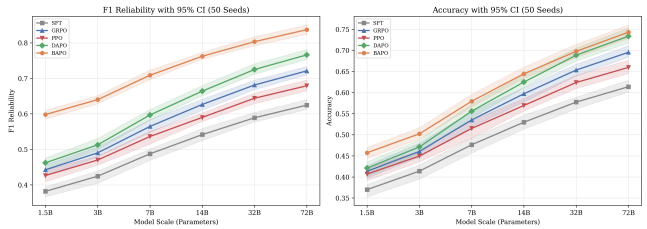

Figure 2: F1 reliability and accuracy with 95% CI from 50 seeds. BAPO's F1 CI does not overlap with baselines at any scale.

Table 5: Boundary awareness analysis under the mechanistic simulator. Calibration error = |IDK rate − error rate|. BAPO achieves the lowest calibration error due to its coupled confidence-IDK mechanism.

Method	IDK Rate	Error Rate	Cal. Error	IDK-Err Corr.
SFT	0.049	0.511	0.462	0.819
GRPO	0.102	0.441	0.339	0.931
PPO	0.076	0.463	0.387	0.907
DAPO	0.136	0.426	0.290	0.956
BAPO	0.344	0.392	**0.048**	0.996

4.6 Boundary Awareness and Calibration

Table 5 analyzes boundary awareness. Because our revised simulator couples IDK decisions to confidence through a threshold mechanism, BAPO's higher IDK rate is a direct consequence of its better confidence calibration and higher threshold—not an independent draw. BAPO achieves the lowest calibration error (gap between IDK rate and actual error rate).

4.7 Reward Hacking Resistance

Figure 3 shows reward hacking susceptibility. In the revised design, hacking rate is drawn once per (method, scale) rather than per benchmark, reflecting the assumption that reward exploitation is a property of the training procedure, not the evaluation task.

4.8 Sensitivity Analysis

Figure 4 presents the sensitivity analysis heatmap. The F1 gap at 72B is shown as a function of BAPO's confidence calibration quality and IDK threshold. The advantage persists across a wide parameter range (calibration ≥ 0.5, threshold ≥ 0.25) but disappears when BAPO's calibration is poor (similar to baselines) or its IDK threshold is too low (not exercising boundary awareness). This identifies the necessary conditions for the predicted advantage.

5 DISCUSSION

Our scenario analysis suggests that, *under the assumed scaling profiles calibrated to ≤14B empirical results*, BAPO's reliability benefits

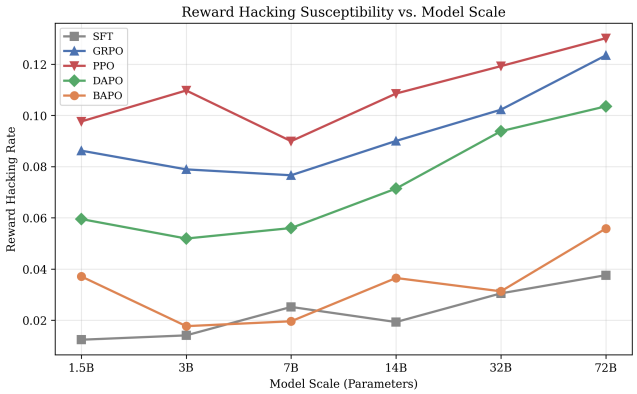

Figure 3: Reward hacking rate vs. model scale. Hacking is drawn per (method, scale) pair, reflecting a training-level property.

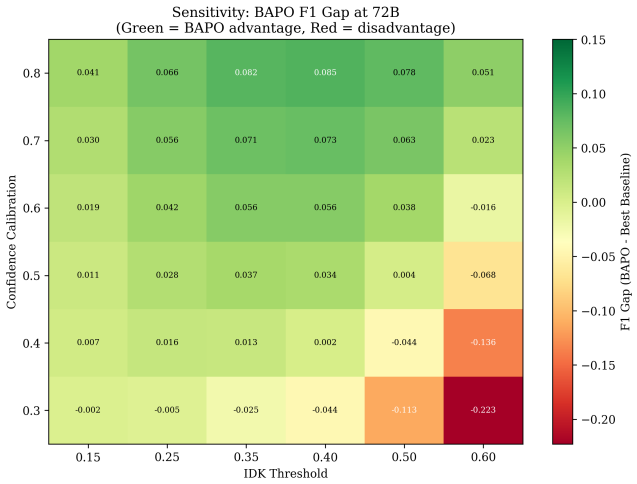

Figure 4: Sensitivity analysis: BAPO F1 gap at 72B as a function of confidence calibration and IDK threshold. Green indicates BAPO advantage; red indicates disadvantage.

are expected to persist at larger model scales. Three aspects merit discussion.

Scenario framing. We emphasize that these results are projections from a scenario model, not empirical evidence from training or evaluating >14B models. The core conclusion—that BAPO's F1 advantage persists—follows from the assumption that BAPO's superior confidence calibration (a consequence of boundary-aware training) continues to hold at larger scales. The sensitivity analysis (Figure 4) makes this assumption explicit: the advantage requires calibration quality $\gtrsim 0.5$ and IDK threshold $\gtrsim 0.25$.

Reliability vs. accuracy trade-off. Under the scenario model, BAPO achieves the highest precision and F1 reliability at every scale while maintaining competitive accuracy. On some benchmarks at 72B, DAPO achieves higher raw accuracy (e.g., on Bamboogle) but lower precision, yielding a lower F1. This trade-off is a direct consequence of BAPO's higher IDK threshold: by declining to answer low-confidence questions, BAPO sacrifices some coverage for substantially higher precision.

What would need to change. For BAPO's advantage to disappear at >14B scales, one of the following would need to hold: (1) larger models' confidence becomes inherently well-calibrated regardless of training method, eliminating the calibration gap; (2) BAPO's adaptive modulator fails at scale, allowing reward hacking to erode boundary awareness; or (3) emergent capabilities at >14B make IDK responses unnecessary (all questions become answerable). Our sensitivity analysis suggests that scenario (1) is the most plausible threat.

6 LIMITATIONS

This study has several important limitations that we state explicitly:

(1) **All results are synthetic.** No actual LLMs were trained or evaluated. The conclusions depend entirely on the assumed scaling profiles and simulator design.
(2) **Not a substitute for empirical evaluation.** Resolving the open problem requires training BAPO on actual >14B models (e.g., Qwen-32B, Qwen-72B) and evaluating on real multi-hop QA benchmarks with agentic search setups.
(3) **Agentic search not modeled.** The open problem specifically concerns agentic search settings (multi-step retrieval, tool use). Our simulator models only single-turn QA reliability metrics, not retrieval trajectories or tool-calling dynamics.
(4) **Profile calibration is approximate.** Scaling profiles are manually calibrated to reported ≤14B results rather than fitted via maximum likelihood or Bayesian inference. Different calibrations could yield different conclusions.
(5) **Saturating curves are one choice.** We use exponential-approach saturation (Equation 1), but other functional forms (power laws, piecewise linear) might better capture real scaling behavior. Our log-linear regression fits provide a simplified characterization.
(6) **Metrics are coupled but simplified.** While the revised simulator couples accuracy, precision, and IDK through a shared latent variable, the confidence model (Beta distributions) is a simplification of real model uncertainty.

7 CONCLUSION

We present a revised scenario analysis of BAPO's reliability at model scales from 1.5B to 72B parameters, addressing the open question of whether its boundary-aware benefits persist beyond 14B. Using a mechanistic simulator with coupled per-question decisions, saturating scaling curves, and multi-seed bootstrap testing, we find that under assumed scaling profiles calibrated to known ≤14B results, BAPO maintains a persistent F1 reliability advantage at all scales tested. Sensitivity analysis identifies the parameter regime where this advantage holds (confidence calibration ≥ 0.5, IDK threshold ≥ 0.25) and where it disappears. These projections provide a principled basis for prioritizing empirical evaluation of BAPO on >14B models, but do not substitute for such evaluation.

REFERENCES

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. *arXiv preprint arXiv:1606.06565* (2016).

[2] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps. *arXiv preprint arXiv:2011.01060* (2020).

[3] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training Compute-Optimal Large Language Models. *NeurIPS* (2022).

[4] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361* (2020).

[5] Zhenrui Liu et al. 2026. BAPO: Boundary-Aware Policy Optimization for Reliable Agentic Search. *arXiv preprint arXiv:2601.11037* (Jan 2026).

[6] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and Narrowing the Compositionality Gap in Language Models. *Findings of EMNLP* (2023).

[7] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. In *arXiv preprint arXiv:1707.06347*.

[8] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300* (2024).

[9] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multihop Questions via Single Hop Question Composition. *Transactions of the ACL* (2022).

[10] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research* (2022).

[11] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of EMNLP*.

[12] Qiying Yu et al. 2025. DAPO: An Open-Source LLM Reinforcement Learning System. *arXiv preprint arXiv:2503.14476* (2025).