

Learning to Think in Blocks: A Prior-Guided Reinforcement Learning Framework for Deep Research

Wangyi Mei¹✦ Ganlin Xu¹✦ Zhoujia Zhang¹✦ Jiaqing Liang¹■†
Weijia Lu²▲ Jinhui Huang¹✦ Chen Yang¹✦ Xiaodong Zhang²▲
Zhifei Yang²▲ Xiaofeng Ma²▲ Deqing Yang¹■†

¹School of Data Science, Fudan University, Shanghai, China

²United Automotive Electronic Systems, Shanghai, China

✦{wymei24, glxu24, zhangzj24, jhhuang25, yang_c25}@m.fudan.edu.cn

■{liangjiaqing, yangdeqing}@fudan.edu.cn

▲{alfredwjlu, xiaodong.zhang.chn, yangzhifei006, maxf0124}@gmail.com

Abstract

Reinforcement learning (RL) for multi-turn deep research agents suffers from a granularity mismatch: token-level methods incur high variance over long reasoning horizons, while sequence-level methods ignore the internal decision structure of multi-turn interactions. We introduce **Block-wise Geometric Policy Optimization (BGPO)**, which treats each assistant turn—packaging reasoning and tool invocation into one semantically coherent unit—as the optimization-level macro-action, and aggregates block-averaged log-importance ratios with equal weights across blocks. We prove that this block-equal weighting minimizes a covariance-robust upper bound on ratio variance under weak inter-block coupling, with the approximation gap relative to sequence-length weighting bounded by block-length imbalance. To improve exploration efficiency, we expose human-defined rewrite operators (decomposition, keyword extraction, synonym replacement, HyDE) as a prompt-level exploration prior during RL rollouts. On seven multi-hop and open-domain QA benchmarks, our 7B system achieves an average judge score of 0.715 ± 0.010 over three training seeds, compared with 0.664 for Search-R1-7B. A controlled fixed-prior attribution study further confirms that block-level optimization independently improves answer quality (0.527/0.395 F1/EM vs. 0.482/0.350 for GRPO). Our code is publicly available on [GitHub](#).

1 Introduction

Deep research systems extend the capability of large language models (LLMs) to handle complex knowledge-intensive tasks through multi-turn retrieval, tool invocation, and intermediate query reformulation (Jin et al., 2025; Song et al., 2025).

Unlike single-pass question answering, these systems must repeatedly decide how to decompose problems, construct queries, invoke tools, and integrate evidence under continuously updated context and external feedback. The key challenge in training deep research agents is therefore not merely generating final answers, but learning a sequence of intermediate research decisions.

Reinforcement learning (RL) provides a natural framework for this goal, as the model can directly optimize multi-turn research strategies through interaction with retrieval and tool environments (Jin et al., 2025; Song et al., 2025). However, existing policy optimization methods exhibit a fundamental *granularity mismatch* with the structure of deep research trajectories. Token-level methods (e.g., PPO, GRPO (Schulman et al., 2017; Shao et al., 2024), DAPO (Yu et al., 2025)) compute importance ratios and clipping at token probabilities, making optimization weights sensitive to tokenization and chunk length rather than corresponding to semantically complete research decisions. Sequence-level methods (e.g., GSPO (Zheng et al., 2025)) assign a single importance ratio to an entire trajectory; while this can improve stability, it conflates query reformulation, tool selection, and evidence integration across different turns into one atomic update unit.

We argue that the natural optimization unit in deep research is the *assistant turn*—a reasoning-tool-use block composed of internal reasoning and a tool invocation. Each block represents a relatively complete research decision: the model forms a next-step research intent based on current context and materializes it as a concrete tool operation. Based on this observation, we propose **Block-wise Geometric Policy Optimization (BGPO)**. BGPO models each assistant turn as a macro-action

†Co-corresponding authors.

$b_k = (\text{think}_k, \text{toolcall}_k)$, computes the average log-importance ratio within each block, and aggregates across blocks with equal weights at the trajectory level. This design avoids token-level length-dominated updates while retaining finer-grained decision resolution than whole-trajectory optimization.

The block-equal aggregation in BGPO does not attempt to recover the exact trajectory importance ratio; instead, it serves as a variance-controlled optimization surrogate. We theoretically characterize the properties of this surrogate: under weak inter-block coupling conditions, block-equal aggregation corresponds to a covariance-robust variance-control criterion; meanwhile, its approximation gap relative to length-weighted sequence aggregation is controlled by block-length imbalance and vanishes when block lengths are uniform. We prove that under weak inter-block coupling—empirically validated in our setting—block-equal aggregation minimizes a covariance-robust upper bound on ratio variance (Theorem 3.1), and that its approximation gap relative to sequence-length weighting remains bounded by block-length imbalance (Proposition 3.2). This analysis explains why assistant-turn-level optimization can simultaneously achieve stability and decision resolution in multi-turn tool-use tasks.

To further improve exploration efficiency in complex research environments, we instantiate BGPO within a prior-guided deep research training framework. During RL rollouts, a set of human-defined query rewrite strategies—including question decomposition, keyword extraction, synonym rewriting, HyDE, and extensible custom strategies—is exposed as optional actions in the prompt. This prompt-level prior guides exploration without changing policy initialization, while the policy learns strategy usage through BGPO under environmental feedback.

Our contributions are: (1) We identify the granularity mismatch between token/sequence-level policy optimization and the decision structure of deep research trajectories, and formalize assistant turns as macro-action blocks. (2) We propose BGPO, which computes block-internal geometric importance ratios with block-equal aggregation, and prove variance-control and bounded approximation-gap guarantees. (3) On seven knowledge-intensive QA benchmarks, BGPO within a prior-guided training framework achieves an average judge score of 0.715 ± 0.010 over three

seeds and the best scale-matched average performance; ablations confirm BGPO’s gain is independent of the prior strategy used.

2 Related Work

Deep Research and Search-Augmented Agents.

Deep research systems extend retrieval-augmented generation (RAG) (Lewis et al., 2020; Guu et al., 2020; Hu and Lu, 2024) to interactive, multi-turn settings where retrieval and reasoning are interleaved through tool use. Search-agent frameworks including Search-o1 (Li et al., 2025), Search-R1 (Jin et al., 2025), R1-Searcher (Song et al., 2025), DeepResearcher (Xu et al., 2025), and the RL-based deep-research system of Wan et al. (2025) train or scaffold agents that interleave reasoning with retrieval. DR-Rubric (Mei et al., 2026) treats rubric construction as a research process and distills retrieved evidence into constraints for policy optimization. Notably, Wan et al. (2025) explicitly define reasoning–tool-use turns, but optimize token actions with a trajectory-level advantage rather than aggregating importance ratios at the turn level. We share this setting but differ in two ways: we provide a prompt-level exploration prior during RL rollouts, and we introduce a block-level optimizer specifically designed for the multi-turn trajectory structure.

Block-Structured Reasoning. Decomposition-style prompting, ReAct (Yao et al., 2023), and IRCot (Trivedi et al., 2023) introduce intermediate reasoning units or search steps, but do not address how importance weights should be distributed across such units during RL. Our contribution is orthogonal: we propose a block-level policy-weighting rule for already block-structured trajectories, rather than a new decomposition scaffold.

Policy Optimization Granularity. Policy optimization for language models has progressed from PPO (Schulman et al., 2017) to GRPO (Shao et al., 2024), DAPO (Yu et al., 2025), and GSPO (Zheng et al., 2025). These methods operate at token level (PPO, GRPO, DAPO) or sequence level (GSPO), but neither granularity matches the natural decision structure of multi-turn reasoning. BGPO introduces an intermediate granularity—the assistant turn—with block-equal importance weighting that bridges token-level variance and sequence-level coarseness, while making the resulting approximation error explicit and bounded.

Concurrent work also studies structure-aware policy-gradient estimation in a different sequential decision domain. ProRL (Hou et al., 2026) shows that positive-mean incremental rewards create a path-extension shortcut in proactive recommendation, while weighting each step by the full path return increases gradient variance; it addresses these issues through stepwise reward centering and position-specific advantage estimation. It is complementary to BGPO: ProRL rectifies reward and advantage estimation over recommendation steps, whereas BGPO restructures importance-ratio aggregation over semantically defined turns in multi-turn language-model agents.

3 Methodology

Notation. We denote the user query as q_{user} , the dialogue state (history and retrieved evidence) at turn k as s_k , and the environment action space as $\mathcal{A}_{\text{env}}$. The policy is parameterized by θ ; $\pi_{\theta}(y_{k,t} | h_{k,t})$ is the token-level probability at position t in block k , where $h_{k,t}$ is the token-level prefix. A block $b_k = (\text{think}_k, \text{toolcall}_k)$ denotes one complete assistant turn; $|b_k|$ is its token length. A trajectory consists of K blocks $\{b_k\}_{k=1}^K$ with total token count $T = \sum_k |b_k|$. The trajectory-level advantage is $\hat{A}_i$ for the i -th sampled trajectory y_i .

3.1 Overview: Three-Level View of Deep Research

Figure 1 illustrates the overall architecture. Our framework separates deep research training into three levels: an environment-facing action space, a block-structured trajectory representation, and a block-level policy optimization objective.

Environment level. The agent interacts with a retrieval environment through a ReAct-style loop (Yao et al., 2023): at each turn it selects an action from a hybrid action space (rewrite, search, or final answer), receives observations, and updates its internal state.

Decision level. A single assistant turn constitutes one complete research decision. Each turn consists of internal reasoning (think) followed by a tool invocation (toolcall). This turn—not individual tokens, not the whole sequence—is the natural unit of research agency.

Optimization level. BGPO does not compute importance ratios at the token level or the whole-sequence level. Instead, it computes a block-level geometric ratio within each block, then aggregates

blocks with equal weights. This ensures every research decision contributes equally to the aggregated log-importance ratio regardless of its verbosity.

3.2 Prior-Guided Research Action Space

Given a user query, the agent repeatedly selects a research action from a hybrid action space consisting of human-prior heuristics, self-devised rewriting, and terminal answering. The human-prior actions provide structured but non-binding exploration biases, while the self-devised action preserves flexibility beyond predefined templates.

Human-Prior Heuristics ($\mathcal{A}_H$). Four curated strategies—*decomposition*, *keyword_extraction*, *synonym_replacement*, and *HyDE* (Gao et al., 2023)—are presented as non-prescriptive descriptions in the system prompt. They provide structured exploration starting points without constraining the policy.

Self-Devised Action ($\mathcal{A}_M$). A flexible strategy *rewrite_by_myself* allows the agent to propose novel queries when predefined heuristics are insufficient.

Terminal Action ($\mathcal{A}_T$). *Generation* terminates the research loop and produces the final answer.

Prompt-Level Exploration Prior. We encode the human-defined rewrite strategies in the system prompt as a prompt-level exploration prior. Their definitions describe how queries may be reformulated, while the model selects a strategy and generates the rewritten query from the current research state. During RL, the policy learns a state-conditioned distribution over the prior-guided and self-devised rewrite actions, adapting strategy selection to the evolving research trajectory.

3.3 Blocks as Optimization Units

Each selected research action is produced through an assistant turn that first reasons about the current dialogue state and then emits a tool invocation. We treat this complete turn as the optimization-level macro-action:

$$b_k = (\text{think}_k, \text{toolcall}_k),$$

where toolcall_k contains the selected environment action $a_k \in \mathcal{A}_{\text{env}}$ and its arguments (e.g., the rewritten query). The think phase represents internal reasoning; the toolcall phase executes a_k and

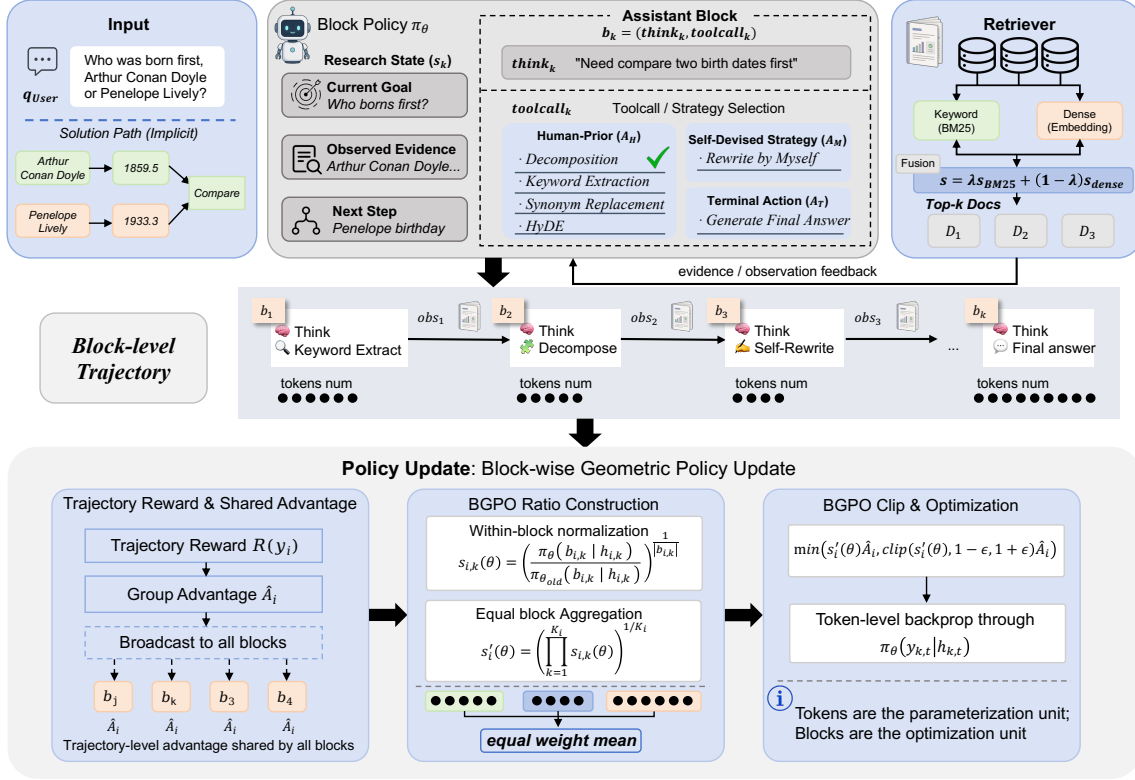


Figure 1: Overview of our framework. The agent interacts with a retrieval environment through block-structured trajectories. Each assistant turn $b_k = (\text{think}_k, \text{toolcall}_k)$ is the optimization unit for BGPO. Human-prior strategies provide a prompt-level exploration prior, and BGPO performs block-level policy optimization with equal weighting across blocks.

induces the state transition to s_{k+1} . A trajectory is thus segmented into blocks $\{b_k\}_{k=1}^K$.

Tokens are the parameterization unit; blocks are the optimization unit. Although the model remains a causal language model generating tokens autoregressively, optimization operates at block granularity. The k -th block, realized by tokens $y_{k,1:|b_k|}$, has probability:

$$\pi_{\theta}(b_k | s_k) = \prod_{t=1}^{|b_k|} \pi_{\theta}(y_{k,t} | h_{k,t}).$$

This factorization connects the block-level optimization abstraction to the underlying token-level parameterization: gradients flow through tokens, but credit is assigned at the block level.

3.4 BGPO: From Token Ratios to Block-Equal Ratios

In deep research, reasoning unfolds through structured assistant turns that interact with the environment. Token-level importance ratios are sensitive to verbosity; sequence-level ratios conflate distinct research decisions. BGPO addresses this by operating at the block level.

For a trajectory y_i with blocks $\{b_{i,k}\}_{k=1}^{K_i}$, we define:

Step 1: Block-level geometric ratio. Using the token-factorized block probability:

$$s_{i,k}(\theta) = \left(\frac{\pi_{\theta}(b_{i,k} | h_{i,k})}{\pi_{\theta_{old}}(b_{i,k} | h_{i,k})} \right)^{\frac{1}{|b_{i,k}|}}, \quad (1)$$

where $h_{i,k} = (b_{i,1}, \dots, b_{i,k-1})$. The $1/|b_{i,k}|$ exponent normalizes within the block, preventing long reasoning chains from producing exploding ratios.

Step 2: Block-equal aggregation. We aggregate across blocks via a geometric mean:

$$s'_i(\theta) = \left(\prod_{k=1}^{K_i} s_{i,k}(\theta) \right)^{\frac{1}{K_i}}, \quad (2)$$

which enforces equal weighting across blocks regardless of their token length.

At the log-ratio level, this corresponds to:

$$\bar{Z}_{\text{blk}}(\theta) = \frac{1}{K} \sum_{k=1}^K \bar{Z}_k(\theta),$$

where $\bar{Z}_k(\theta) := \frac{1}{|b_k|} \sum_{t=1}^{|b_k|} \log \frac{\pi_{\theta}(y_{k,t} | h_{k,t})}{\pi_{\theta_{old}}(y_{k,t} | h_{k,t})}$, in contrast to the conventional sequence-length

weighting $\bar{Z}_{\text{seq}}(\theta) = \sum_{k=1}^K \frac{|b_k|}{T} \bar{Z}_k(\theta)$ with $T = \sum_k |b_k|$.

Although the resulting surrogate $s'_i(\theta)$ is a trajectory-level scalar, it differs from sequence-level importance weighting because its log form assigns equal weight to each assistant turn rather than to each token.

Step 3: BGPO objective. The final objective accumulates gradients at the token level while sharing the block-structured importance ratio:

$$\mathcal{J}_{\text{BGPO}}(\theta) = \mathbb{E}_{\substack{x \sim \mathcal{D}, \\ \{y_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}}} \left[\frac{1}{\sum_j |y_j|} \sum_{i=1}^G \sum_{k=1}^{K_i} \sum_{t=1}^{|b_{i,k}|} \min \left(s'_i(\theta) \hat{A}_i, \text{clip}(s'_i(\theta), 1-\epsilon, 1+\epsilon) \hat{A}_i \right) \right], \quad (3)$$

where $\hat{A}_i$ is a trajectory-level advantage shared across all tokens.

To summarize, the three-level normalization structure is: (i) *block-internal normalization* via the $1/|b_k|$ exponent avoids ratio explosion from long blocks; (ii) *block-equal aggregation* via the $1/K$ exponent prevents long blocks from dominating the trajectory-level update; (iii) *token-level gradient accumulation*—actual backpropagation still flows through token parameters.

3.5 Why Equal Blocks? Variance-Bias Tradeoff

The block-equal surrogate deliberately departs from exact sequence likelihood correction. We therefore analyze two questions: whether equal block weighting improves stability, and how much approximation error it introduces relative to length-weighted sequence aggregation. Full proofs are in Appendix B.

Theorem 3.1 (Variance minimization under weak inter-block coupling). Consider a trajectory with $K \geq 2$ blocks. Define the block-averaged log-importance ratio:

$$\bar{Z}_k(\theta) := \frac{1}{|b_k|} \sum_{t=1}^{|b_k|} \log \frac{\pi_{\theta}(y_{k,t} | h_{k,t})}{\pi_{\theta_{\text{old}}}(y_{k,t} | h_{k,t})}. \quad (4)$$

Assume $\mathbb{E}[\bar{Z}_k(\theta)] = 0$, $\text{Var}(\bar{Z}_k(\theta)) = \tau_k^2 \leq \tau_{\text{max}}^2$, and weak inter-block coupling:

$$\left| \sum_{k \neq \ell} w_k w_{\ell} \text{Cov}(\bar{Z}_k(\theta), \bar{Z}_{\ell}(\theta)) \right| \leq \lambda \sum_{k=1}^K w_k^2 \tau_k^2 \quad (5)$$

for every $w \in \Delta_K$. For any convex combination $\bar{Z}_w(\theta) = \sum_k w_k \bar{Z}_k(\theta)$:

$$\text{Var}(\bar{Z}_w(\theta)) \leq (1 + \lambda) \tau_{\text{max}}^2 \sum_{k=1}^K w_k^2. \quad (6)$$

This upper bound is minimized by $w_k^{\text{blk}} = 1/K$, giving $(1 + \lambda) \tau_{\text{max}}^2 / K$. Sequence-length weights $w_k^{\text{seq}} = |b_k| / \sum_j |b_j|$ yield a strictly larger bound whenever block lengths differ.

Proposition 3.2 (Surrogate bias controlled by block-length dispersion). Define $\bar{Z}_{\text{blk}}(\theta)$ and $\bar{Z}_{\text{seq}}(\theta)$ as above. If $|\bar{Z}_k(\theta)| \leq C$ almost surely, then

$$|\bar{Z}_{\text{blk}}(\theta) - \bar{Z}_{\text{seq}}(\theta)| \leq C \sum_{k=1}^K \left| \frac{1}{K} - \frac{|b_k|}{T} \right|. \quad (7)$$

The gap vanishes when block lengths are uniform and grows with block-length dispersion.

Theorem 3.1 answers the stability question: under weak inter-block coupling—empirically validated in our setting—block-equal weights minimize the covariance-robust variance upper bound. Proposition 3.2 answers the bias question: the approximation gap is bounded by block-length imbalance and vanishes when blocks are similar in length. Together, they characterize BGPO as trading exact importance correction for a block-equal surrogate with better variance control, at the cost of a bounded and empirically small approximation gap. The practical implication is that variance-controlled updates may improve policy-gradient stability and convergence; Section 4 examines whether the observed training dynamics are consistent with this prediction.

4 Experiments

4.1 Setup

We evaluate on seven QA benchmarks spanning multi-hop reasoning (HotpotQA (Yang et al., 2018), 2WikiMultihopQA (Ho et al., 2020), MuSiQue (Trivedi et al., 2022)), open-domain (NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017)), and search-heavy (Bamboogle (Press et al., 2023), PopQA (Mallen et al., 2023)) settings. Training uses filtered splits of HotpotQA/2Wiki/MuSiQue; all test numbers are on original dev sets. Evaluation uses an LLM-as-Judge protocol (Qwen3-32B); the attribution study

additionally reports F1/EM to remove judge dependence. Unless otherwise stated, results are from a single run; for BGPO-7B in Table 1, we report the mean and sample standard deviation over three training seeds. Baseline implementation provenance and retriever settings are summarized in Appendix A.6, Table 8.

Harder benchmarks. We additionally evaluate on three harder deep-research benchmarks that demand longer search trajectories: **BrowseComp+** (Chen et al., 2025) is a controlled variant of BrowseComp (Wei et al., 2025) that provides a fixed, human-verified corpus with supporting documents and challenging negatives, enabling fair comparison without relying on live web access; **HoVer** (Jiang et al., 2020) is a many-hop fact extraction and claim verification dataset requiring evidence from up to four Wikipedia articles with diverse reasoning graph structures; **SealQA** (Pham et al., 2025) is an adversarially curated benchmark where web search yields conflicting, noisy, or unhelpful results, testing whether models can reason under retrieval uncertainty.

Harder baselines. Beyond the main baselines (Direct, CoT, Naive RAG, IRCoT, Search-o1, Search-R1), we compare against two stronger 7B search-agent baselines on the harder benchmarks: **SimpleDeepSearcher** (Sun et al., 2025) synthesizes high-quality reasoning trajectories via web-powered data generation and achieves strong deep-search performance through supervised fine-tuning on only 871 curated samples; **ASearcher** (Gao et al., 2025) employs fully asynchronous reinforcement learning at scale to train agents capable of long-horizon (up to 128 turns) agentic search with uncertainty-aware reasoning and cross-document inference.

Full dataset statistics, hyperparameters, and baseline descriptions are in Appendices A.1–A.

4.2 Overall Results

Table 1 presents the full benchmark comparison. We summarize results by task category:

Multi-hop reasoning. BGPO-7B obtains 0.877 ± 0.006 on HotpotQA (+0.067 over Search-R1-7B), 0.799 ± 0.016 on 2Wiki (+0.076), and 0.701 ± 0.057 on MuSiQue (+0.063). The larger variation on MuSiQue cautions against overinterpreting its margin.

Open-domain and search-heavy. BGPO-7B improves over Search-R1-7B on NQ (+0.037), TriviaQA (+0.047), and PopQA (+0.072), while Bamboogle is within 0.002.

Scale comparison. At 7B, BGPO averages 0.715 ± 0.010 , compared with 0.664 for Search-R1-7B. At 14B, BGPO scores 0.767 versus 0.699 for Search-R1-14B. DR-Rubric scores 0.585 using its public Qwen3-8B checkpoint and is reported as a public-checkpoint comparison because its backbone and agent design differ.

Cross-backbone generalization. We further test whether BGPO transfers beyond the Qwen family by repeating the comparison on Llama-3.1-8B-Instruct over the same seven benchmarks.

Table 2 shows that BGPO achieves the best average score (0.635 vs. 0.618 for IRCoT) and leads on six of seven datasets; IRCoT remains best on PopQA. This cross-family result indicates that the block-level formulation is not tied to a Qwen backbone.

4.3 Controlled Attribution: Optimizer vs. Prior

The main benchmark mixes optimizer choice, prior-guided training, and system-level effects. We isolate each factor below.

4.3.1 Fixed Prior, Vary Optimizer

We fix the prior-guided rollout protocol and compare four optimizers—GRPO, GSPO, DAPO, and BGPO—under matched conditions, evaluated with judge-independent F1/EM.

Table 3 shows that Prior+BGPO achieves the best overall average (0.527 F1 / 0.395 EM), outperforming Prior+DAPO (0.498 / 0.363), Prior+GRPO (0.482 / 0.350), and Prior+GSPO (0.443 / 0.307). BGPO exceeds GSPO on all seven datasets under both metrics and leads the complete optimizer set on six of seven datasets. Since the prior is held fixed and evaluation uses F1/EM, this result isolates the optimizer’s independent contribution to answer quality.

Retrieval-depth diagnostic. Under the same prior-guided action space and training configuration, GSPO averages 1.5–1.9 retrieval rounds across the seven benchmarks, versus approximately 3.9 for BGPO. On the 1,000-question HotpotQA evaluation, all GSPO trajectories terminate within two rounds, while 93% of BGPO trajectories reach

Method	Scale	Multi-hop*			Open-domain [†]		Search-heavy [‡]		Avg.
		HotpotQA*	2Wiki*	MuSiQue*	NQ [†]	TriviaQA [†]	Bamboogle [‡]	PopQA [‡]	
Prompt / RAG baselines									
Direct Answer	7B	0.298	0.252	0.111	0.382	0.397	0.567	0.371	0.339
CoT	7B	0.322	0.305	0.142	0.392	0.424	0.520	0.298	0.343
NaiveRAG	7B	0.424	0.400	0.240	0.610	0.433	0.016	0.037	0.308
Search methods									
IRCoT	7B	0.540	0.705	0.484	0.660	0.488	0.416	0.208	0.500
IRCoT	14B	0.851	0.779	0.583	0.698	0.703	0.560	0.482	0.665
Search-o1	7B	0.526	0.694	0.396	0.554	0.673	0.440	0.263	0.506
Search-o1	14B	0.780	0.549	0.580	0.649	0.682	0.502	0.435	0.608
Search-R1	7B	0.810	0.723	0.638	0.755	0.702	0.547	0.472	0.664
Search-R1	14B	0.873	0.792	0.654	0.759	0.797	<u>0.584</u>	0.437	0.699
DR-Rubric	8B	0.679	0.615	0.446	0.717	0.695	0.511	0.433	0.585
Our methods									
Qwen2.5 (BGPO)	7B	<u>0.877</u> ^{±0.006}	<u>0.799</u> ^{±0.016}	<u>0.701</u> ^{±0.057}	<u>0.792</u> ^{±0.016}	0.749 ^{±0.007}	0.545 ^{±0.036}	<u>0.544</u> ^{±0.036}	<u>0.715</u> ^{±0.010}
Qwen2.5 (BGPO)	14B	0.923	0.867	0.802	0.814	<u>0.796</u>	0.604	0.563	0.767

Table 1: Main judge scores on multi-hop, open-domain, and search-heavy QA benchmarks. BGPO-7B reports means with sample standard deviations shown as superscripts over three training seeds; other rows are single runs. Best mean performance is bold and second-best is underlined. */[†]/[‡] denote in-domain, out-of-domain, and search-heavy settings. Full baseline implementation provenance is reported in Table 8.

Baseline	HotpotQA	2Wiki	MuSiQue	NQ	PopQA	TriviaQA	Bamboogle	Avg.
NaiveRAG	0.812	0.584	0.474	0.623	0.435	0.611	0.259	0.542
Search o1	0.638	0.562	0.409	0.560	0.403	0.580	0.460	0.516
IRCoT	0.847	0.720	0.524	0.628	0.463	0.643	0.499	0.618
Ours (BGPO)	0.853	0.758	0.561	0.640	0.431	0.664	0.539	0.635

Table 2: Cross-backbone generalization on Llama-3.1-8B-Instruct. BGPO achieves the best average and leads on six of the seven datasets.

four rounds. This contrast shows earlier answering under GSPO and sustained multi-round exploration under BGPO.

4.3.2 Fixed Optimizer, Remove Prior

Table 4 isolates the prior’s contribution under a fixed BGPO optimizer: (1) BGPO without prior already outperforms prior-only reasoning, confirming that the optimization algorithm provides gains independent of the prompt-level prior. (2) The full system yields the largest improvements (+0.055 on HotpotQA, +0.082 on 2Wiki, +0.120 on MuSiQue over prior-only), indicating that BGPO exploits prior-induced trajectory structure. (3) Prior and BGPO are complementary: neither alone matches the combined system.

4.4 Mechanistic Evidence: Stable BGPO Optimization Dynamics

The controlled attribution establishes *what* BGPO achieves; we now examine *how* the four prior-guided runs behave over their shared 1,596-step window.

Figure 2 visualizes four optimization metrics

over the shared 1,596-step window. BGPO reaches its sustained 95%-plateau in 666 steps, earlier than GSPO (988), DAPO (860), and GRPO (854). BGPO and GSPO both maintain near-zero mean importance-ratio standard deviation (0.006 and 0.005) with zero clipping, whereas DAPO and GRPO exhibit higher standard deviation (0.127 and 0.964) and nonzero mean clip fractions (0.026 and 0.024). BGPO also achieves the highest mean batch reward (1.519) and longest mean output length (2005) among the four methods, consistent with stable updates supporting deeper exploration.

These dynamics are consistent with the theoretical prediction that block-level ratio decomposition controls update variability; the downstream F1/EM comparisons provide separate evidence for answer quality.

4.5 Theoretical Assumption Validation

Our theory relies on two key assumptions: weak inter-block coupling (Theorem 3.1) and bounded surrogate bias (Proposition 3.2). We validate both empirically.

Method	HotpotQA F1/EM	2Wiki F1/EM	MuSiQue F1/EM	NQ F1/EM	TriviaQA F1/EM	Bamboogle F1/EM	PopQA F1/EM	Avg. F1/EM
Prior+GRPO	0.692 / 0.548	0.688 / 0.583	0.516 / 0.385	0.497 / 0.376	0.322 / 0.137	0.455 / 0.336	0.206 / 0.086	0.482 / 0.350
Prior+GSPO	0.672 / 0.503	0.621 / 0.503	0.484 / 0.323	0.486 / 0.353	0.294 / 0.110	0.349 / 0.272	0.195 / 0.082	0.443 / 0.307
Prior+DAPO	0.709 / 0.551	0.705 / 0.594	0.553 / 0.414	0.527 / 0.408	0.324 / 0.128	0.467 / 0.368	0.198 / 0.081	0.498 / 0.363
Prior+BGPO	0.740 / 0.589	0.716 / 0.607	0.619 / 0.483	0.563 / 0.448	0.332 / 0.134	0.514 / 0.407	0.205 / 0.094	0.527 / 0.395

Table 3: Optimizer attribution: F1/EM under an identical prior-guided action space. Each cell is F1/exact-match, intentionally judge-independent. BGPO achieves the best average on both metrics and leads on 6/7 datasets.

Method	HotpotQA	2Wiki	MuSiQue
Prior Only	0.778	0.691	0.467
BGPO w/o prior	0.784	0.725	0.494
Ours (BGPO)	0.833	0.773	0.587

Table 4: Effect of the prompt-level exploration prior under a fixed BGPO optimizer. “BGPO w/o prior” trains without the prior-guided action descriptions.

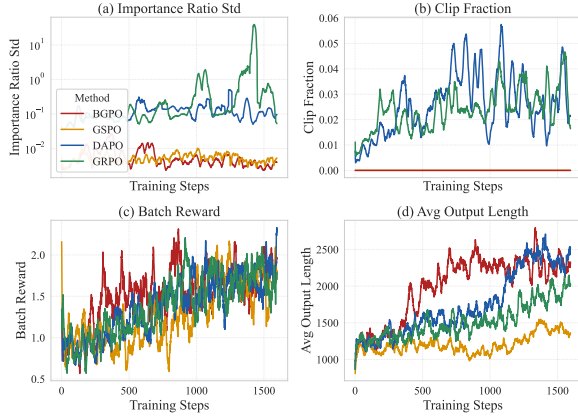


Figure 2: Training dynamics over the shared first 1,596 steps of four prior-guided runs. BGPO (red) and GSPO (orange) maintain near-zero importance-ratio standard deviation with zero clipping, while BGPO achieves the highest mean batch reward and longest mean output length. Panel (a) uses a logarithmic vertical axis.

Weak coupling. Figure 3 shows that inter-block correlations remain small at all training stages (mean $|\hat{\rho}| = 0.14\text{--}0.18$, max < 0.24), validating the independence assumption in Theorem 3.1.

Surrogate bias. The empirical gap $|\bar{Z}_{\text{blk}} - \bar{Z}_{\text{seq}}|$ stays at $1.1\text{--}1.3 \times 10^{-2}$ across checkpoints (P90 < 0.029), confirming that the block-equal weighting approximation cost is small in practice (Proposition 3.2).

Narrative closure. These diagnostics close the loop: theory predicts variance control $\rightarrow$ diagnostics validate assumptions $\rightarrow$ BGPO exhibits lower ratio standard deviation and zero clipping than the token-level baselines $\rightarrow$ the controlled F1/EM com-

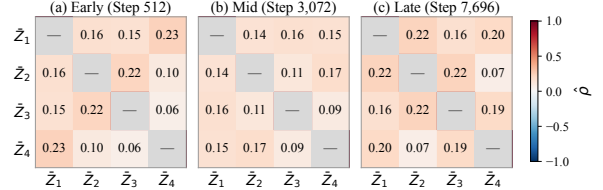


Figure 3: Pairwise correlation $\hat{\rho}_{kl}$ of block-level advantage fluctuations at three training stages. All off-diagonal entries satisfy $|\hat{\rho}_{kl}| < 0.24$, confirming weak inter-block coupling throughout training.

parison shows higher answer quality. Full details are in Appendices E–E.1.

4.6 Harder Deep-Research Benchmarks

Table 6 evaluates on three harder deep-research datasets (BrowseComp+, HoVer, SealQA) that require substantially longer search trajectories (average 5–8 retrieval rounds vs. 2–3 for the main seven benchmarks), making them a stress test for optimizer stability. BGPO leads on all three with an average of 0.242 vs. 0.224 for Search-R1. The margin is most visible on HoVer (0.629 vs. 0.588), where multi-hop fact verification demands sustained reasoning over many retrieved passages; gains on BrowseComp+ and SealQA are smaller but directionally consistent. We use this slice to support a narrower claim than raw leaderboard superiority: the block-level structure is most justified in longer-horizon search settings where shallow rewriting policies are less reliable.

Judge Robustness. A 200-sample blind human evaluation agrees with the Qwen3-32B judge (Cohen’s $\kappa = 0.688$), and a Llama-3.3-70B judge agrees on 88.1% of 12,250 rescored predictions. Since the attribution result in Table 3 is anchored by F1/EM, it does not depend on judge choice.

4.7 Case Study

Table 5 highlights three failure modes that BGPO avoids: (1) *semantic confusion*—ablated variants retrieve correct evidence but misinterpret entity

Dataset	Query	BGPO (RL + Prior)	BGPO w/o Prior (RL only)	Prior Only (no RL)
HotpotQA	Which road borders both Pacific Mall and Splendid China Mall? <i>Gold: Steeles Avenue</i>	Strategy: decomposition Queries: “location Pacific Mall”; “location Splendid China Mall”; “common road bordering both” Reasoning: Pacific Mall is on Steeles Ave & Kennedy Rd; Splendid China Mall is at Redlea Ave & Steeles Ave. Common border = Steeles Avenue. Answer: Steeles Avenue ✓	Strategy: rewrite_by_myself Query: “What are the locations of Pacific Mall and Splendid China Mall?” Reasoning: Retrieves correct info but confuses entrance road (Redlea Ave) with border road. Answer: Redlea Avenue ✗	Strategy: keyword_extraction Query: “road, Pacific Mall, Splendid China Mall” Reasoning: Same confusion between entrance and border roads. Answer: Redlea Avenue ✗
2Wiki	Who died first, Earl Valiquette or Jean-Joseph Gaume? <i>Gold: Jean-Joseph Gaume</i>	Strategy: keyword_extraction Queries: “Earl Valiquette death”; “Jean-Joseph Gaume death” Reasoning: Gaume died 1879, Valiquette died 1975. 1879 < 1975, so Gaume died first. Answer: Jean-Joseph Gaume ✓	Strategy: rewrite_by_myself Query: “Earl Valiquette death date and Jean-Joseph Gaume death date” Reasoning: Retrieves same dates but reverses the comparison: concludes Valiquette died first. Answer: Earl Valiquette ✗	Strategy: keyword_extraction Query: “Earl Valiquette death date Jean-Joseph Gaume death date” Reasoning: Same reversal error in temporal comparison. Answer: Earl Valiquette ✗
MuSiQue	Who is the employer of the physicist who Mach number is named after? <i>Gold: Charles University</i>	Strategy: decomposition → keyword → rewrite (4 rounds) Process: (1) identifies Ernst Mach; (2–3) searches employer without direct hit; (4) infers from Dvořák’s assistant record at Prague. Answer: Charles University ✓	Strategy: rewrite_by_myself (2 rounds) Process: Two direct queries for “employer of Ernst Mach” fail. Falls back to parametric guess. Answer: University of Vienna ✗	Strategy: keyword_extraction (2 rounds) Process: First query retrieves irrelevant “Full-time employment” article. Falls back to guess. Answer: University of Zagreb ✗

Table 5: Case study on multi-hop QA. **Green** = correct key reasoning; **Red** = error point. BGPO leverages strategy diversity and RL-trained execution to succeed where ablated variants fail.

Method	BrowseComp+	HoVer	SealQA	Avg.
Search-R1	0.0609	0.5880	0.0219	0.224
simple_deep_searcher	0.0482	0.5150	0.0165	0.193
ASearcher	0.0444	0.4760	0.0291	0.183
Qwen2.5 (BGPO)	0.0630	0.6290	0.0353	0.242

Table 6: Performance on harder deep-research benchmarks requiring longer search trajectories. BGPO leads on all three datasets.

roles (entrance vs. border road); (2) *reasoning reversal*—given identical dates, ablated models invert temporal comparisons; (3) *early abandonment*—without RL-trained persistence, models give up after two retrieval rounds and resort to parametric guessing. BGPO overcomes these via decomposition-based queries and adaptive multi-round strategy switching, consistent with the quantitative gains in Table 4.

Additional analyses. Training-set-size sensitivity and extended case studies are in Appendices C.1 and D; retrieval robustness (top- k , sparse/dense weight) and real-world web search (+79.5% on SimpleQA) are in Appendices E.2–E.3.

5 Conclusion

We introduced Block-wise Geometric Policy Optimization (BGPO), which resolves the granularity mismatch between existing RL methods and multi-turn deep research by treating each assistant turn as the optimization unit. BGPO’s block-equal

weighting minimizes a covariance-robust variance upper bound under weak inter-block coupling, with a surrogate bias bounded by block-length dispersion. Combined with a prompt-level exploration prior, our framework achieves an average judge score of 0.715 ± 0.010 over three seeds and the best scale-matched average performance on seven QA benchmarks; controlled attribution confirms that BGPO independently improves optimization stability over token-level baselines.

Limitations

Our study is subject to several limitations as follows. First, the manually designed rewriting priors may not generalize to all multi-turn reasoning tasks. Second, BGPO defines blocks over the ReAct-style format; asynchronous, nested, or multi-tool interactions require an adapted block schema. Third, our evaluation focuses on retrieval-augmented QA, primarily with Qwen backbones and automatic judges, although we additionally conduct a 200-sample blind human evaluation; broader model families, scales, languages, agent domains, and larger-scale human evaluation remain open. Finally, our theory has an assumption of bounded block heterogeneity and weak inter-block coupling, which may not always hold in all multi-turn reasoning tasks.

Ethical Considerations

Our work focuses on improving the reasoning and information-seeking capabilities of deep research

agents. While such systems can facilitate access to information, they may retrieve or synthesize inaccurate, outdated, biased, or otherwise unreliable content. Improving their capabilities may also amplify these risks when the generated reports are used without verification. Our method does not directly address source reliability, privacy, or potential misuse. Therefore, practical deployments should incorporate source attribution, factual verification, appropriate access controls, and human oversight, particularly in high-stakes domains.

Acknowledgments

This work was supported by the AI Laboratory of United Automotive Electronic Systems (UAES) Co. and the Chinese NSF General Program (No. 62572129).

References

- Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [BGE m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). *CoRR*, abs/2402.03216.
- Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, Sahel Sharifmoghaddam, Yanxi Li, Haoran Hong, Xinyu Shi, Xuye Liu, Nandan Thakur, Crystina Zhang, Luyu Gao, Wenhui Chen, and Jimmy Lin. 2025. [Browsecomp-plus: A more fair and transparent evaluation benchmark of deep-research agent](#). *CoRR*, abs/2508.06600.
- Jiaxuan Gao, Wei Fu, Minyang Xie, Shusheng Xu, Chuyi He, Zhiyu Mei, Banghua Zhu, and Yi Wu. 2025. [Beyond ten turns: Unlocking long-horizon agentic search with large-scale asynchronous RL](#). *CoRR*, abs/2508.07976.
- Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. [Precise zero-shot dense retrieval without relevance labels](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2023, Toronto, Canada, July 9-14, 2023, pages 1762–1777. Association for Computational Linguistics.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. [Retrieval augmented language model pre-training](#). In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.
- Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. [Constructing A multi-hop QA dataset for comprehensive evaluation of reasoning steps](#). In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 6609–6625. International Committee on Computational Linguistics.
- Hongru Hou, Tiehua Mei, Denghui Geng, Jinhui Huang, Ao Xu, Hengrui Chen, Jiaqing Liang, and Deqing Yang. 2026. [ProRL: Effective reinforcement learning for proactive recommendation via rectified policy gradient estimation](#). *Preprint*, arXiv:2605.28293.
- Yucheng Hu and Yuxing Lu. 2024. [RAG and RAU: A survey on retrieval-augmented language model in natural language processing](#). *CoRR*, abs/2404.19543.
- Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. [HoVer: A dataset for many-hop fact extraction and claim verification](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online, November 16-20, 2020*, pages 3441–3460. Association for Computational Linguistics.
- Bowen Jin, Hansi Zeng, Zhenrui Yue, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. [Search-r1: Training llms to reason and leverage search engines with reinforcement learning](#). *CoRR*, abs/2503.09516.
- Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. [Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. [Natural questions: a benchmark for question answering research](#). *Trans. Assoc. Comput. Linguistics*, 7:452–466.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). In *Advances in Neural Information Processing Systems 33: Annual Conference*

- ence on Neural Information Processing Systems 2020, *NeurIPS 2020, December 6-12, 2020, virtual*.
- Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025. [Search-ol: Agentic search-enhanced large reasoning models](#). *CoRR*, abs/2501.05366.
- Jiaqing Liang. 2025. [Lsrll: Memory efficient large model training framework](#). <https://github.com/ldefine/lsrll>.
- Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khoshnab, and Hannaneh Hajishirzi. 2023. [When not to trust language models: Investigating effectiveness of parametric and non-parametric memories](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2023, Toronto, Canada, July 9-14, 2023, pages 9802–9822. Association for Computational Linguistics.
- Wangyi Mei, Zhouhong Gu, Zhenhan Bai, Yin Cai, Lefan Zhang, Zhenxin Ding, Bo Chen, Yan Gao, Yi Wu, Yao Hu, Jiaqing Liang, and Deqing Yang. 2026. [Deep research as rubric for reinforcement learning](#). *CoRR*, abs/2606.01091.
- Thinh Pham, Nguyen Nguyen, Pratibha Zunjare, Weiyuan Chen, Yu-Min Tseng, and Tu Vu. 2025. [Sealqa: Raising the bar for reasoning in search-augmented language models](#). *CoRR*, abs/2506.01062.
- Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. [Measuring and narrowing the compositionality gap in language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 5687–5711. Association for Computational Linguistics.
- Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. [Okapi at TREC-3](#). In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, volume 500-225 of *NIST Special Publication*, pages 109–126. National Institute of Standards and Technology (NIST).
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. [Proximal policy optimization algorithms](#). *CoRR*, abs/1707.06347.
- 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](#). *CoRR*, abs/2402.03300.
- Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025. [R1-searcher: Incentivizing the search capability in llms via reinforcement learning](#). *CoRR*, abs/2503.05592.
- Shuang Sun, Huatong Song, Yuhao Wang, Ruiyang Ren, Jinhao Jiang, Junjie Zhang, Fei Bai, Jia Deng, Wayne Xin Zhao, Zheng Liu, Lei Fang, Zhongyuan Wang, and Ji-Rong Wen. 2025. [Simpledeepsearcher: Deep information seeking via web-powered reasoning trajectory synthesis](#). *CoRR*, abs/2505.16834.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. [Musique: Multi-hop questions via single-hop question composition](#). *Trans. Assoc. Comput. Linguistics*, 10:539–554.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. [Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2023, Toronto, Canada, July 9-14, 2023, pages 10014–10037. Association for Computational Linguistics.
- Yi Wan, Jiuqi Wang, Liam Li, Jinsong Liu, Ruihao Zhu, and Zheqing Zhu. 2025. [Rethinking the design of reinforcement learning-based deep research agents](#). *CoRR*, abs/2510.15862.
- Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. [Measuring short-form factuality in large language models](#). *CoRR*, abs/2411.04368.
- Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. 2025. [Browsecomp: A simple yet challenging benchmark for browsing agents](#). *CoRR*, abs/2504.12516.
- Guowei Xu, Jinxi Li, Yichen Jiang, Xupei Huang, Haopeng Chen, Ruizhe Zhang, Xianhao Zeng, Zhengyuan Yang, Lu Qi, Kaisheng Zeng, Yimin Dou, Yasheng Wang, Zhengdian Liu, Hao Zhou, Zhoujun Li, and Jiuming Liu. 2025. [Deepresearcher: Scaling deep research via reinforcement learning in real-world environments](#). *CoRR*, abs/2504.03160.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 40 others. 2025. [Qwen3 technical report](#). *CoRR*, abs/2505.09388.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. [Qwen2.5 technical report](#). *CoRR*, abs/2412.15115.
- 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 the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. [React: Synergizing reasoning and acting in language models](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, and 16 others. 2025. [DAPO: an open-source LLM reinforcement learning system at scale](#). *CoRR*, abs/2503.14476.
- Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. 2025. [Group sequence policy optimization](#). *CoRR*, abs/2507.18071.

Set	Dataset	# Samples
Train Set	HotpotQA	2,075
	2WikiMultihopQA	1,731
	MuSiQue	1,325
Test Set	HotpotQA	7,405
	2WikiMultihopQA	12,576
	MuSiQue	2,417
	NQ	6,515
	TriviaQA	8,837
	PopQA	3,567
	Bamboogle	125

Table 7: Statistics of the datasets used for training and evaluation.

A Additional Experimental Details

A.1 Datasets, Splits, and Filtering

We use HotpotQA (Yang et al., 2018), 2WikiMultihopQA (Ho et al., 2020), MuSiQue (Trivedi et al., 2022), NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), PopQA (Mallen et al., 2023), and Bamboogle (Press et al., 2023). We form the training mixture using filtered training splits of HotpotQA, 2WikiMultihopQA, and MuSiQue, and evaluate on disjoint, original dev splits. We call these three benchmarks in-domain because their training splits contribute to training. NQ and TriviaQA are out-of-domain because none of their training examples are used, while PopQA and Bamboogle are likewise unseen during training and are reported separately as search-heavy benchmarks. Table 7 reports dataset statistics. For the follow-up generalization slices, we additionally evaluate on BrowseComp+, HoVer, and SealQA as harder search benchmarks in Table 6, and on a separate Llama-3.1-8B-Instruct backbone in Table 2.

Filtering protocol. For the training set, we filter out noisy instances (e.g., ambiguous reasoning chains or missing evidence) to reduce spurious reward signals and mitigate reward hacking. We preserve data difficulty and diversity: in the retained HotpotQA training subset, 23.95% of examples are labeled *hard*, and 25.35% require reasoning over 3+ supporting facts. Importantly, all results in Table 1 are evaluated on the original, unfiltered test/dev sets, ensuring performance is measured over the full difficulty distribution.

Corpora and retrieval sources. For NQ and TriviaQA, we use the Wikipedia corpus released by Karpukhin et al. (2020). For PopQA and Bamboogle, we query Wikipedia via online search rather than a static local corpus. To further reflect real-world retrieval complexity, we additionally evaluate on SimpleQA using a real-time internet search engine API.

A.2 Evaluation Protocol

We adopt an LLM-as-Judge evaluation protocol. The judge model is Qwen3-32B (Yang et al., 2025), constrained to output only one token from {Yes, No}. Given a question, a reference answer, and a model prediction, we compute the score as the probability assigned to Yes at the first generated token. The full evaluation prompt is provided in Appendix A.3.

A.3 Prompt for evaluating the answer

Prompt.

You are a strict AI evaluator. Your sole task is to compare two answers and determine if they agree.

- First, analyze the core meaning of Answer A. - Next, analyze the core meaning of Answer B. - Finally, decide if their fundamental conclusions are the same. - Focus on pure factual agreement, ignoring any stylistic or minor differences.

You MUST conclude your final response with ONLY ONE WORD on a new line: 'Yes' or 'No'. Do not provide any other text or explanation.

Question: {question}

Answer A: {ground_truth}

Answer B: {model_answer}

Here is an example:

You are a strict AI evaluator. Your sole task is to compare two answers and determine if they agree.

- First, analyze the core meaning of Answer A. - Next, analyze the core meaning of Answer B. - Finally, decide if their fundamental conclusions are the same. - Focus on pure factual agreement, ignoring any stylistic or minor differences.

You MUST conclude your final response with ONLY ONE WORD on a new line: 'Yes' or 'No'. Do not provide any other text or explanation.

Question: Were Scott Derrickson and Ed Wood of the same nationality?

Answer A: yes

Answer B: Scott Derrickson is American, and Ed Wood is also American. Therefore, Scott Derrickson and Ed Wood were of the same nationality.

A.4 System prompt for the RAG system

Prompt.

- **Your Role:** You are a search strategist for a Retrieval-Augmented Generation (RAG) system. Your **ONLY** job is to decide the best way to search for information in an external knowledge base.
- **CRITICAL RULE:** You **MUST NOT** use your own internal knowledge to answer the question. Assume you know nothing. Your entire reasoning process must be about how to find the information, not what the information is. Even if you know the answer, your task is to formulate search queries to find supporting documents.
- **The Goal:** To intelligently select or create search queries (rewrite strategies) in a step-by-step process to gather all necessary facts from the knowledge base to answer the user’s query.
- **User’s Query:** “{query}”
- **Available Strategies:** You can use the following strategies to formulate the next search query. Rewrite strategies:
 1. decomposition: Break down the original query into several simpler sub-queries for better retrieval.
 2. keyword_extraction: Extract key terms from the query to search in order to focus the retrieval on some specific terms.
 3. synonym_replacement: Replace some words in the query with their synonyms to enable the retrieval to search beyond the original terms.
 4. HyDE: Generate hypothetical answers as query to bridge the gap between the expression of query and the expression of docs in the knowledge base.
 5. rewrite_by_myself: Rewrite the query adaptively based on the previous context and your own understanding of the query.
- **When to Stop:** Only choose to “generate the final answer” when you believe the information that would be hypothetically retrieved using your sequence of queries is sufficient to answer the user’s question. Your decision to stop must be based on the completeness of your search plan, not on your internal knowledge of the answer.

A.5 Training and System Configuration

We implement training with the Low-Resource Reinforcement Learning (LSRL) package (Liang, 2025) for low-resource, high-throughput RL with gradient offloading. Our backbone model is Qwen2.5-7B-Instruct (Yang et al., 2024). Training is run on 4×A800 GPUs for approximately 2 days.

Retriever. We use a hybrid retriever combining sparse and dense retrieval. Sparse retrieval is BM25 (Robertson et al., 1994) and dense retrieval is BGE-m3 (Chen et al., 2024). We set the sparse/dense mixture weight to 0.5 and retrieve top-3 documents. All baselines and our method share the same retriever and corpus.

Optimization hyperparameters. We train for 3 epochs. For each training instance, we sample 8 trajectories. We use learning rate $1e-6$ with training batch size 8 and gradient accumulation over 32 steps. The KL penalty coefficient is $\beta = 0.001$, and the PPO-style upper clipping parameter is 0.28. For rollout generation, we use batch size 32, temperature 0.9, and cap tokens per round at 768; the maximum context length is 5120 tokens. We enable CPU gradient offloading to reduce memory usage.

Resource allocation. We use 1 GPU for the reward model, 1 GPU for vLLM rollouts, and 2 GPUs for data-parallel training.

A.6 Baselines

We compare against the following baselines.

Direct Answer. The model produces an answer directly without exposing intermediate reasoning.

Chain-of-Thought (CoT). The model first generates step-by-step reasoning and then outputs the final answer.

Naive RAG. The model performs a single retrieval step to collect relevant documents and generates an answer conditioned on the retrieved context.

IRCoT (Trivedi et al., 2023). Retrieval is interleaved with CoT reasoning steps, allowing intermediate reasoning to guide subsequent retrieval and vice versa.

Search-o1 (Li et al., 2025). The model assesses retrieved information for relevance and veracity before incorporating it into the reasoning process.

Search-R1 (Jin et al., 2025). A reinforcement learning framework that combines reasoning and retrieval. In the main comparison, we report 7B baselines throughout and additionally include 14B variants for IRCoT, Search-o1, and Search-R1. Unless otherwise stated, all methods share the same retriever and corpus for fair comparison.

simple_deep_searcher and ASearcher. These are stronger 7B search-agent baselines used in the harder-benchmark comparison of Table 6. We keep their original agent designs and compare them under the same judge protocol as the rest of the paper.

Method	Backbone / checkpoint	Size	Agent loop	Retriever	Impl.
Direct / CoT	Qwen2.5-Instruct	7B	–	Same	Ours
NaiveRAG	Qwen2.5-Instruct	7B	Single-step	Same	Ours
IRCoT	Qwen2.5-Instruct	7B, 14B	ReAct-style	Same	Reimpl.
Search-o1	Qwen2.5-Instruct	7B, 14B	ReAct-style	Same	Reimpl.
Search-R1	Qwen2.5-Instruct	7B, 14B	ReAct-style	Same	Reimpl.
DR-Rubric	public Qwen3 RL-bootstrap-step 3	8B	ReAct-style	Same	Public
Prior+GRPO/DAPO/GSPO/BGPO	Qwen2.5-7B-Instruct	7B	ReAct-style	Same	Ours

Table 8: Implementation provenance of the main baselines. Agent loop: Single-step performs one retrieval, while ReAct-style interleaves reasoning and retrieval. Same: the same hybrid retriever (BM25+BGE-m3, top-3, mixture weight 0.5); Ours: our implementation; Reimpl.: our reimplementation; Public: the DR-Rubric public release.

DR-Rubric reproduction boundary. We first attempted a Qwen2.5-7B reproduction of the DR-Rubric training recipe, but the resulting model did not learn active retrieval and obtained $F1 = 0.086$. The main comparison therefore uses the closest available public checkpoint, Qwen3-8B RL-bootstrap-step 3, retains its native ReAct-style agent design, routes its search calls to our retriever, and evaluates it with the same Qwen3-32B judge. Because the backbone, checkpoint, and agent protocol differ, this row is not a controlled optimizer comparison; Table 3 provides that comparison.

B Full Theoretical Proofs

B.1 Proof of Turn-Level Stability in BGPO

We now formalize the stability property induced by the block-equal aggregation mechanism used in BGPO. The result shows that, under bounded turn-wise heteroscedasticity and weak inter-block coupling, equal weighting across reasoning blocks minimizes a covariance-robust worst-case upper bound on the variance of the aggregated log-importance ratio.

Theorem B.1 (Approximate stability of block-equal aggregation under weak inter-block coupling). Consider a trajectory composed of $K \geq 2$ assistant reasoning blocks (turns). For each block $k \in \{1, \dots, K\}$, we define the block-averaged log-importance ratio as

$$\bar{Z}_k(\theta) := \frac{1}{|b_k|} \sum_{t=1}^{|b_k|} \log \frac{\pi_\theta(y_{k,t} | h_{k,t})}{\pi_{\theta_{\text{old}}}(y_{k,t} | h_{k,t})}. \quad (8)$$

Assume that under the rollout distribution induced by $\pi_{\theta_{\text{old}}}$ and the environment,

$$\mathbb{E}[\bar{Z}_k(\theta)] = 0, \quad \text{Var}(\bar{Z}_k(\theta)) = \tau_k^2 \leq \tau_{\max}^2. \quad (9)$$

Instead of assuming exact block independence, suppose the inter-block dependence is weak in the fol-

lowing sense: there exists a constant $\lambda \geq 0$ such that for every $w \in \Delta_K$,

$$\left| \sum_{k \neq \ell} w_k w_\ell \text{Cov}(\bar{Z}_k(\theta), \bar{Z}_\ell(\theta)) \right| \leq \lambda \sum_{k=1}^K w_k^2 \tau_k^2. \quad (10)$$

For any convex weights $w \in \Delta_K := \{w \in \mathbb{R}_+^K : \sum_{k=1}^K w_k = 1\}$, we define the aggregated log-ratio estimator as

$$\bar{Z}_w(\theta) := \sum_{k=1}^K w_k \bar{Z}_k(\theta). \quad (11)$$

Then the following hold:

1. Robust variance bound.

$$\begin{aligned} \text{Var}(\bar{Z}_w(\theta)) &\leq (1 + \lambda) \sum_{k=1}^K w_k^2 \tau_k^2 \\ &\leq (1 + \lambda) \tau_{\max}^2 \sum_{k=1}^K w_k^2. \end{aligned} \quad (12)$$

2. Minimax optimality for the worst-case upper bound.

Let $w_k^{\text{blk}} = 1/K$. Then

$$\begin{aligned} \inf_{w \in \Delta_K} (1 + \lambda) \tau_{\max}^2 \sum_{k=1}^K w_k^2 \\ = (1 + \lambda) \tau_{\max}^2 \sum_{k=1}^K (w_k^{\text{blk}})^2 = \frac{(1 + \lambda) \tau_{\max}^2}{K}. \end{aligned} \quad (13)$$

3. Comparison with sequence-length weighting.

Let $w_k^{\text{seq}} := \frac{|b_k|}{\sum_{j=1}^K |b_j|}$. Then

$$(1 + \lambda) \tau_{\max}^2 \sum_{k=1}^K (w_k^{\text{seq}})^2 \geq \frac{(1 + \lambda) \tau_{\max}^2}{K}, \quad (14)$$

with strict inequality whenever the block lengths are not identical.

Therefore, under bounded turn-wise heteroscedasticity and weak inter-block coupling, block-equal aggregation minimizes a covariance-robust worst-case upper bound on the variance of the aggregated log-importance ratio. In the special case $\lambda = 0$, the result reduces to the exact independence statement.

Proof. By variance decomposition,

$$\begin{aligned} \text{Var}(\bar{Z}_w(\theta)) &= \sum_{k=1}^K w_k^2 \tau_k^2 \\ &\quad + \sum_{k \neq \ell} w_k w_\ell \text{Cov}(\bar{Z}_k(\theta), \bar{Z}_\ell(\theta)). \end{aligned} \quad (15)$$

Applying (5) gives

$$\begin{aligned} \text{Var}(\bar{Z}_w(\theta)) &\leq \sum_{k=1}^K w_k^2 \tau_k^2 + \lambda \sum_{k=1}^K w_k^2 \tau_k^2 \\ &= (1 + \lambda) \sum_{k=1}^K w_k^2 \tau_k^2 \\ &\leq (1 + \lambda) \tau_{\max}^2 \sum_{k=1}^K w_k^2, \end{aligned} \quad (16)$$

which proves part (1).

It remains to minimize $\sum_{k=1}^K w_k^2$ over $w \in \Delta_K$. By Cauchy–Schwarz,

$$\left(\sum_{k=1}^K w_k \right)^2 \leq \left(\sum_{k=1}^K 1^2 \right) \left(\sum_{k=1}^K w_k^2 \right), \quad (17)$$

and since $\sum_{k=1}^K w_k = 1$, this implies

$$\sum_{k=1}^K w_k^2 \geq \frac{1}{K}, \quad (18)$$

with equality if and only if $w_k = 1/K$ for all k . Substituting this into (16) gives

$$\inf_{w \in \Delta_K} (1 + \lambda) \tau_{\max}^2 \sum_{k=1}^K w_k^2 = \frac{(1 + \lambda) \tau_{\max}^2}{K}, \quad (19)$$

establishing (13) and proving part (2). Applying the same inequality to w^{seq} yields (14), with strict inequality whenever the block lengths are not identical. This proves part (3). $\square$

B.2 Surrogate Bias Relative to Sequence-Length Weighting

Theorem 3.1 explains the variance-control benefit of block-equal aggregation. We now make explicit what objective is being approximated when BGPO replaces sequence-length weighting by block-equal weighting.

Proposition B.2 (Length-dispersion control of surrogate bias). Define

$$\begin{aligned} \bar{Z}_{\text{blk}}(\theta) &:= \frac{1}{K} \sum_{k=1}^K \bar{Z}_k(\theta), \\ \bar{Z}_{\text{seq}}(\theta) &:= \sum_{k=1}^K \frac{|b_k|}{T} \bar{Z}_k(\theta), \end{aligned} \quad (20)$$

where $T = \sum_{k=1}^K |b_k|$, and let

$$\begin{aligned} \Delta(\theta) &:= \bar{Z}_{\text{blk}}(\theta) - \bar{Z}_{\text{seq}}(\theta) \\ &= \sum_{k=1}^K \left(\frac{1}{K} - \frac{|b_k|}{T} \right) \bar{Z}_k(\theta). \end{aligned} \quad (21)$$

Assume $|\bar{Z}_k(\theta)| \leq C$ almost surely for all k . Then

$$|\Delta(\theta)| \leq C \sum_{k=1}^K \left| \frac{1}{K} - \frac{|b_k|}{T} \right|. \quad (22)$$

If in addition the normalized block lengths satisfy

$$\left| \frac{|b_k|}{T} - \frac{1}{K} \right| \leq \frac{\delta}{K} \quad \text{for all } k, \quad (23)$$

then

$$|\Delta(\theta)| \leq C\delta. \quad (24)$$

More generally, for any L -Lipschitz surrogate map ℓ on an interval containing both $\bar{Z}_{\text{blk}}(\theta)$ and $\bar{Z}_{\text{seq}}(\theta)$,

$$\begin{aligned} &|\mathbb{E}[\ell(\bar{Z}_{\text{blk}}(\theta))] - \mathbb{E}[\ell(\bar{Z}_{\text{seq}}(\theta))]| \\ &\leq LC \sum_{k=1}^K \left| \frac{1}{K} - \frac{|b_k|}{T} \right|. \end{aligned} \quad (25)$$

Finally, if $\bar{Z}_k(\theta)$ are centered independent σ -sub-Gaussian random variables, then $\Delta(\theta)$ is centered sub-Gaussian with variance proxy

$$\sigma_{\Delta}^2 = \sigma^2 \sum_{k=1}^K \left(\frac{1}{K} - \frac{|b_k|}{T} \right)^2, \quad (26)$$

so that for every $\varepsilon > 0$,

$$\Pr(|\Delta(\theta)| \geq \varepsilon) \leq 2 \exp\left(-\frac{\varepsilon^2}{2\sigma_\Delta^2}\right). \quad (27)$$

Therefore, BGPO approximates the sequence-length-weighted objective with a bias that vanishes for uniform block lengths and increases with block-length dispersion.

Proof. Equation (21) is obtained by direct subtraction of the two aggregators. Applying the triangle inequality and the bound $|\bar{Z}_k(\theta)| \leq C$ yields

$$\begin{aligned} |\Delta(\theta)| &\leq \sum_{k=1}^K \left| \frac{1}{K} - \frac{|b_k|}{T} \right| \cdot |\bar{Z}_k(\theta)| \\ &\leq C \sum_{k=1}^K \left| \frac{1}{K} - \frac{|b_k|}{T} \right|, \end{aligned} \quad (28)$$

which proves (7). Under (23), each summand is at most δ/K , so

$$|\Delta(\theta)| \leq C \sum_{k=1}^K \frac{\delta}{K} = C\delta, \quad (29)$$

proving (24).

For any L -Lipschitz ℓ ,

$$|\ell(\bar{Z}_{\text{blk}}(\theta)) - \ell(\bar{Z}_{\text{seq}}(\theta))| \leq L|\Delta(\theta)|. \quad (30)$$

Taking expectations and using (7) gives (25).

Under the centered independent σ -sub-Gaussian assumption, any weighted sum remains sub-Gaussian with variance proxy equal to the squared weights times σ^2 . Applying this standard closure property to (21) yields the stated expression for σ_Δ^2 and the tail bound (27). $\square$

B.3 Connection to Policy Gradient Stability

We now clarify how Theorem 3.1 and Proposition 3.2 should be interpreted in relation to policy-gradient stability. Theorem 3.1 is a weak-coupling stability result rather than an exact empirical law. When inter-block covariance vanishes, it recovers the exact minimax variance statement. In the empirically observed regime, inter-block correlations remain small but non-zero, so the theorem is best viewed as showing that block-equal weighting minimizes a covariance-robust upper bound on ratio variance. Proposition 3.2 complements this by showing that the induced surrogate bias is small when block lengths are not too dispersed. Together, they justify BGPO as a variance-control surrogate consistent with the observed training regime.

Equivalence of Implicit Weighting Schemes

Consider the log-ratio contributions $z_{k,t} := \log \frac{\pi_\theta(y_{k,t}|h_{k,t})}{\pi_{\theta_{\text{old}}}(y_{k,t}|h_{k,t})}$.

Token-level PPO/GRPO and sequence-level GSPO both use importance weights of the form $\exp(\frac{1}{T} \sum_{k,t} z_{k,t})$, where $T = \sum_k |b_k|$. This is equivalent to a block-weighted aggregation

$$\bar{Z}_{\text{seq}}(\theta) = \sum_{k=1}^K w_k^{\text{seq}} \bar{Z}_k(\theta), \quad w_k^{\text{seq}} = \frac{|b_k|}{\sum_j |b_j|}, \quad (31)$$

i.e., token-length weighting across blocks.

In contrast, BGPO uses the hierarchical geometric ratio $\exp(\frac{1}{K} \sum_k \bar{Z}_k)$, corresponding to block-equal weights $w_k^{\text{blk}} = 1/K$.

Therefore, Theorem 3.1 compares the covariance-robust upper bounds induced by sequence-length weighting and block-equal weighting, rather than claiming exact dominance of realized variance in every empirical setting.

From Ratio Variance to Gradient Stability Let the one-sample policy gradient estimator be

$$\hat{g}(\theta) = A \cdot s(\theta) \cdot \nabla_\theta \log \pi_\theta(y | x), \quad (32)$$

where $s(\theta) = \exp(\bar{Z}_w(\theta))$ is the aggregated importance ratio and A is the advantage.

Assume:

1. $|A| \leq A_{\max}$,
2. $\|\nabla_\theta \log \pi_\theta(y | x)\| \leq G_{\max}$,
3. $|\bar{Z}_w(\theta)| \leq C$ (e.g., via clipping or bounded log-ratio).

Then, $s(\theta) = e^{\bar{Z}_w}$ is Lipschitz in $\bar{Z}_w$ with constant $L = e^C$, and

$$\|\hat{g}(\theta)\| \leq A_{\max} G_{\max} e^C. \quad (33)$$

By the Lipschitz continuity of the estimator with respect to $\bar{Z}_w$, we have

$$\text{Var}(\hat{g}(\theta)) \leq (A_{\max} G_{\max} e^C)^2 \text{Var}(\bar{Z}_w(\theta)). \quad (34)$$

Combining this with Theorem 3.1 yields

$$\begin{aligned} \text{Var}(\hat{g}(\theta)) &\leq (A_{\max} G_{\max} e^C)^2 \\ &\quad \cdot (1 + \lambda) \tau_{\max}^2 \sum_{k=1}^K w_k^2. \end{aligned} \quad (35)$$

Hence, block-equal weighting minimizes a covariance-robust worst-case upper bound on one-step gradient variance under weak inter-block coupling. This supports the observed stability trend of BGPO rather than asserting an exact minimax empirical law.

Dataset	1k (F1/EM)	3k (F1/EM)	5k (F1/EM)	10k (F1/EM)
HotpotQA	0.674 / 0.533	0.668 / 0.499	<u>0.680 / 0.540</u>	0.731 / 0.574
2Wiki	0.628 / 0.536	0.625 / 0.516	<u>0.692 / 0.595</u>	0.750 / 0.647
MuSiQue	<u>0.523 / 0.404</u>	0.507 / 0.377	0.488 / 0.384	0.581 / 0.474
NQ	0.514 / 0.405	0.510 / 0.387	<u>0.522 / 0.407</u>	0.547 / 0.428
TriviaQA	0.325 / 0.143	0.317 / 0.125	0.304 / 0.116	<u>0.321 / 0.133</u>
Bamboogle	0.425 / 0.320	0.336 / 0.256	0.366 / 0.272	<u>0.383 / 0.272</u>
PopQA	0.183 / <u>0.084</u>	<u>0.190 / 0.082</u>	0.171 / 0.075	0.209 / 0.096
Avg.	<u>0.467 / 0.346</u>	0.450 / 0.320	0.460 / 0.341	0.503 / 0.375

Table 9: Train-size sensitivity report for BGPO. Each cell is reported as F1 / EM, with all seven benchmark rows taken from the matched 1k/3k/5k/10k evaluation suite. The average row is computed over all seven datasets. Best values are in **bold**; second-best values are underlined.

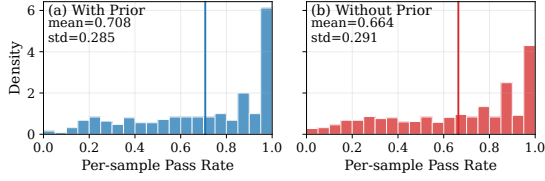


Figure 4: Distribution of per-sample rollout pass rates with and without the prompt-level exploration prior.

C Additional Experiments on Generalization and Robustness

This appendix collects secondary evidence complementing the main-text attribution and robustness analyses, including train-size sensitivity, harder deep-research benchmarks, cross-backbone transfer, a qualitative case study, weak-coupling diagnostics, and real-world retrieval robustness checks.

C.1 Training-Set-Size Sensitivity of BGPO

We evaluate data-scale sensitivity using matched 1k/3k/5k/10k training subsets across all seven benchmark datasets (HotpotQA, 2Wiki, MuSiQue, NQ, TriviaQA, Bamboogle, and PopQA). The resulting pattern is mixed at intermediate scales but consistently strongest at 10k on four of the seven datasets, so we report the evidence conservatively without claiming a strictly monotonic trend at every size.

C.2 Prior-Guided Exploration Diagnostics

The prior-guided model achieves higher median accuracy (~ 0.80 vs. ~ 0.70) with slightly lower dispersion, indicating that the prompt-level prior improves exploration during rollout sampling.

D Additional Case Study

In Figure 5, we present a qualitative analysis using a query about the “founder of India’s civil services.”

Case Study: Multi-Step Decomposition and Final Generation	
Question:	Who introduced the system of civil services in India?
Observation Snippet:	The agent retrieves text that suggests a nuanced answer with multiple contributors, requiring careful synthesis. <i>...Warren Hastings laid the foundation of civil service and Charles Cornwallis reformed, modernised and rationalised it. Hence, Charles Cornwallis is known as the 'Father of Civil Service in India'.</i>
Agent's Process:	Intermediate Step (Decomposition): Initially, the agent breaks the ambiguous query into simpler sub-questions to gather foundational knowledge. • When was the civil service system introduced in India? • Who established the civil service in India? Final Answer (Synthesis): After retrieving and processing information, the agent synthesizes the details into a precise final answer, correctly attributing the main role. [ANSWER] Charles Cornwallis is known as the “Father of Civil Service in India” and he introduced Covenanted Civil Services (Higher Civil Services) and Uncovenanted Civil Services (Lower Civil Services).

Figure 5: Case study of the model using human priors to solve a complex RAG problem.

Metric	Early (512)	Mid (3072)	Late (7696)
Mean $ \hat{\rho}_{k\ell} $	0.14	0.16	0.18
Max $ \hat{\rho}_{k\ell} $	0.21	0.22	0.24
Inflation Ratio	1.41	1.46	1.52

Table 10: Summary statistics of weak inter-block coupling analysis across training stages.

The retrieved documents contain conflicting information mentioning two different historical figures, which makes direct answering difficult.

Guided by the learned strategy, the agent resolves this ambiguity through a structured reasoning procedure. It first triggers the **Decomposition** action to break the question into fact-based sub-queries, enabling targeted evidence collection for each candidate. During the subsequent **Synthesis** stage, the agent integrates the verified evidence to correctly attribute the role to Charles Cornwallis. This example illustrates that the policy does not merely retrieve supporting text, but learns a multi-step reasoning workflow for handling ambiguity.

E Detailed Weak Inter-Block Coupling Analysis

This section provides additional empirical evidence for the weak inter-block coupling regime used in Theorem 3.1 (see Section 4 for the main-text summary). As shown in Figure 3, all off-diagonal correlations satisfy $|\hat{\rho}_{k\ell}| < 0.24$ throughout training. Below we provide detailed summary statistics of the covariance structure of block-level advantage fluctuations $\bar{Z}_k(\theta)$ at three training checkpoints: Early (step 512), Mid (step 3072), and Late (step 7696).

Table 10 summarizes the key statistics. The slight increase in correlation from Early to Late training is expected, as the policy becomes more structured over time, leading to marginally stronger

Step	Pairs	Disp.	Mean $ \Delta $	P90 $ \Delta $	V_b/V_s
512	500	0.214	0.0115	0.0252	0.964
3072	410	0.199	0.0125	0.0286	0.957
7696	465	0.192	0.0107	0.0247	0.965

Table 11: Empirical audit of the surrogate bias in Proposition 3.2. The gap remains small throughout training, and the block-equal variance proxy stays close to the sequence-weighted one.

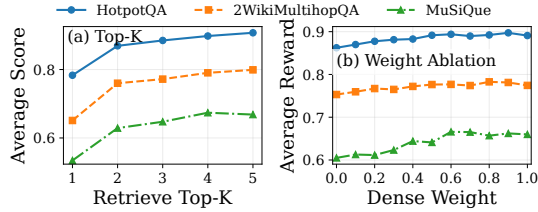


Figure 6: Ablation study on retrieval parameters.

inter-block dependencies. Nevertheless, even at the latest checkpoint, the inflation ratio remains well below 2.0, indicating that inter-block covariance acts as a moderate inflation term rather than contradicting the robust upper-bound interpretation of Theorem 3.1.

E.1 Empirical Surrogate-Bias Audit

To complement Proposition 3.2, we directly measure the empirical gap $\Delta(\theta) = \bar{Z}_{\text{blk}}(\theta) - \bar{Z}_{\text{seq}}(\theta)$ on the same saved covariance checkpoints used for the weak-coupling audit. For each checkpoint, we pair the saved $\{\bar{Z}_k\}$ values with the corresponding rollout trajectory, reconstruct assistant-turn block lengths from the stored messages, and compute both the block-equal and sequence-length-weighted log-ratio aggregators sample by sample.

The absolute gap remains concentrated near zero throughout training: the mean $|\Delta|$ stays around 1.1×10^{-2} to 1.3×10^{-2} , and even the 90th percentile never exceeds 0.029. At the same time, the reconstructed block-length dispersion is modest (mean L1 dispersion 0.192–0.214), matching the theoretical prediction that the approximation error grows with block-length heterogeneity. The last column shows that the empirical variance of the block-equal proxy remains within about 4% of the sequence-length-weighted proxy at all three checkpoints. We therefore use this audit to support a narrow interpretation of Proposition 3.2: the BGPO surrogate is not exact, but on the saved training trajectories its bias is small and empirically traceable.

Method	NQ	TriviaQA	SimpleQA
NaiveRAG	0.599	0.698	0.215
Deep Research + BGPO	0.692	0.737	0.386

Table 12: Performance under real-time web search within an Iterative Deep Research system. Incorporating BGPO yields consistent gains across datasets, indicating improved robustness to noisy, open-web retrieval. This suggests that the prior-guided + BGPO stack remains effective under imperfect retrieval signals, while not by itself isolating optimizer-only effects.

E.2 Retriever Parameter Robustness

Increasing the number of retrieved documents from $\text{top-}k = 1$ to $\text{top-}k = 3$ yields substantial gains, with diminishing returns thereafter. Hybrid sparse-dense retrieval remains strong across a range of mixture ratios, supporting the shared $\text{top-}k = 3$, weight-0.5 setting used in the main comparisons.

E.3 Robustness in Real-World Web Search

To validate the practical applicability of our framework in a realistic, noisy environment, we integrated a real-time internet search engine API. We evaluated the model on the **SimpleQA** (Wei et al., 2024) benchmark (measuring short-form factuality) as well as the online versions of **NQ** and **TriviaQA**.

As shown in Table 12, RAG-BGPO significantly outperforms NaiveRAG even when dealing with the complexity of live web search results. Notably, on SimpleQA, our method achieves a **79.5% relative improvement** (0.386 vs 0.215). This demonstrates that our iterative rewriting strategies effectively filter the noise of the open web and generalize well to short-form factual queries, without overfitting to the long-context offline datasets used during training.