Can David Beat Goliath? ON MULTI-HOP REASONING WITH RESOURCE-CONSTRAINED AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning (RL) trains small language model agents to answer multi-hop questions by retrieving evidence over multiple turns, but reported gains typically rely on thousands of on-policy rollouts per update. We study RL for such agents under the budget constraint of commodity GPUs, where each update samples only a few rollouts per question. Under this constraint, most sampled trajectories retrieve none of the required evidence, so the outcome reward gives the policy little to learn from and small agents settle for answering without retrieval, a failure we call *retrieval collapse*. DAVID-GRPO addresses this with two mechanisms: (1) *Expert trajectory seeding* places a handful of off-policy expert trajectories into the GRPO groups of the early updates, and (2) *evidence-guided continuation* rewards evidence coverage and resumes the most promising partial trajectory. The evidence for each training question is constructed from the corpus link graph, so no annotated evidence is required. On six multi-hop QA benchmarks, DAVID-GRPO trained on four RTX 3090 GPUs with 144 rollouts per step brings Qwen2.5-1.5B to 22.6 average EM against 11.9 for the best baseline under the same budget, matches Tree-GRPO trained with 20 times more rollouts, and, unlike the baselines that stop after at most one search, learns to retrieve across turns.[1]

1 INTRODUCTION

Reinforcement learning (RL) has recently been used to train small language model agents that interleave reasoning with retrieval on long-horizon tasks such as multi-hop question answering (Jin et al., 2025b; Ji et al., 2025; Zhang et al., 2025; Ji et al., 2026; Chen et al., 2025). The reported gains, however, come from high-cost training with thousands of on-policy rollouts per update and large batches (Figure 1), which is out of reach for most groups working on commodity hardware (Kudiabor, 2024; Besiroglu et al., 2024).

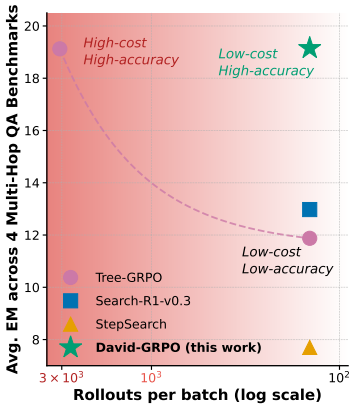


Figure 1: Average EM on four multi-hop QA benchmarks versus rollouts per batch (log scale, Qwen2.5-1.5B). Shading marks rollout intensity and the dashed line traces Tree-GRPO scaling. With 4.7% of the high-budget rollouts, DAVID-GRPO matches high-budget Tree-GRPO.

In contrast, we study commodity GPU scenarios with a small rollout budget, where RL fails in a specific way. GRPO-style training forms advantages from reward differences among the trajectories sampled for the same question (Shao et al., 2024). In multi-hop retrieval, a trajectory is useful only if it retrieves the bridge evidence and then the answer evidence, and with few rollouts most sampled trajectories retrieve none of it. The outcome reward then gives the policy little to learn from, and small agents settle for answering from parametric knowledge without retrieval. We call this failure *retrieval collapse*. With no rewarded retrieval in a group, GRPO gives no positive signal for multi-turn retrieval and reinforces answering without it. Every prior low-budget agent we train undergoes retrieval collapse and converges to zero

[1] https://anonymous.4open.science/r/David-GRPO-5261/

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

or one search per question (§5). Existing ways to add learning signal assume resources that a low rollout budget lacks. SFT initialization requires thousands of expert demonstrations (Zhang et al., 2025), tree expansion requires large batches to find useful prefixes (Ji et al., 2026), and retrieval-level rewards require annotated evidence for every training question (Jin et al., 2025a; Zheng et al., 2025).

We introduce DAVID-GRPO, which learns multi-turn retrieval under a low rollout budget with two mechanisms. Expert trajectory seeding prevents retrieval collapse by placing a handful of off-policy expert trajectories into the GRPO groups of the early updates, so that each group contains rewarded multi-turn retrieval before the agent's own rollouts retrieve evidence (§3.2). Evidence-guided continuation rewards the fraction of evidence a trajectory retrieves and resumes the best partial trajectory from the last successful retrieval, so that the agent extends partial retrieval chains until the evidence is complete (§3.3). The evidence for each training question is constructed from the corpus link graph (§4.1), so no annotated evidence is required.

On six multi-hop QA benchmarks, DAVID-GRPO trained on four RTX 3090 GPUs with 144 rollouts per step and seeded with four expert trajectories brings Qwen2.5-1.5B to 22.6 average EM against 11.9 for the best baseline under the same budget, and matches Tree-GRPO trained with 20 times more rollouts (Table 1). The improvement over the baselines holds on the Qwen2.5 and Llama backbones we test and, in a smaller experiment, on a 7B agent (§4.2). Unlike the baselines, which stop after at most one search, DAVID-GRPO learns to retrieve across turns (§5).

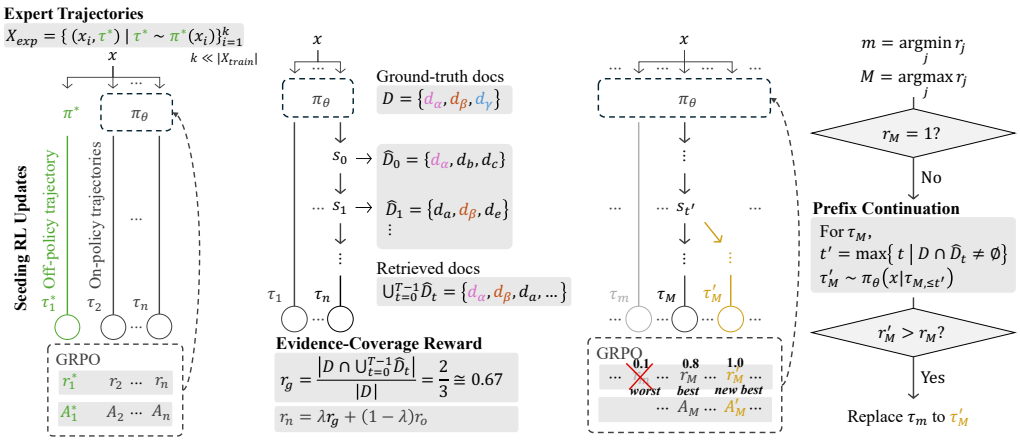
2 RELATED WORK

Search Agents Trained with RL. Multi-hop QA requires an agent to retrieve several documents from a corpus and combine them (Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022; Press et al., 2023). Retrieval serves two roles here, *bridge retrieval* for intermediate entities and *answer retrieval* for the final evidence (Xiong et al., 2019; Fang et al., 2020; Xiong et al., 2021). Recent work trains such agents with RL, mostly from outcome rewards on the final answer (Jin et al., 2025b; Zhang et al., 2025; Chen et al., 2025; Ji et al., 2026), using thousands of on-policy rollouts per update. With small rollout batches, few trajectories earn a reward and these agents settle for answering after zero or a single search (§5).

Expert Demonstrations and Rollout Expansion. When an update contains few useful trajectories from the current policy, one approach distills trajectories from a stronger model through SFT before RL, as AutoCoA does with thousands of multi-turn agent trajectories (Zhang et al., 2025). Another is to reuse partially successful on-policy trajectories. Tree-GRPO randomly expands a reasoning tree and propagates outcome rewards to intermediate steps (Ji et al., 2026). Neither fits a low budget. The SFT stage needs thousands of demonstrations, and random expansion chooses expansion points without evidence signals and so needs large batches to hit useful prefixes by chance.

Retrieval Rewards and Evidence Annotation. Retrieval-level rewards give denser feedback than outcome rewards. StepSearch scores each step by lexical similarity to the annotated evidence (Zheng et al., 2025), and Search-R1-v0.3 rewards retrieving the answer document (Jin et al., 2025a). Lexical similarity is a noisy proxy for retrieving the evidence itself, both rewards credit individual documents and can be satisfied by a single hit, and both need annotated evidence for every training question. Existing training sets obtain this evidence from crowdworkers, who write a question over a chain of documents and mark the supporting paragraphs (Yang et al., 2018; Trivedi et al., 2022). We follow the same chain-based construction but let a language model write the questions, as in synthetic QA generation (Alberti et al., 2019), so the chain itself serves as the evidence.

Our Distinction. DAVID-GRPO needs neither large annotation sets nor large batches. Expert trajectory seeding puts a handful of expert trajectories (four in our experiments, §4.1) into the GRPO groups of the early updates in place of an SFT stage (§3.2). Evidence-guided continuation rewards the fraction of the required evidence that a trajectory has retrieved, so credit grows only as bridge and answer evidence accumulate (§3.3.1), and when no trajectory in a group reaches the maximum reward, prefix continuation resumes the highest-coverage trajectory at the step where coverage last increased (§3.3.2). The evidence and the questions come from the corpus graph without human supervision. In our experiments, DAVID-GRPO matches or exceeds the best low-budget baseline on every backbone, matches high-budget Tree-GRPO, and learns multi-turn retrieval.



(a) Expert Trajectory Seeding. (b) Evidence-Guided Continuation.

Figure 2: Overview of DAVID-GRPO.

3 METHOD

We first formalize the task (§3.1), then present the two components of DAVID-GRPO, expert trajectory seeding (§3.2) and evidence-guided continuation (§3.3). Figure 2 overviews DAVID-GRPO with these components.

3.1 PROBLEM FORMULATION

We follow the retrieval-based multi-hop QA setting, where the information needed to answer a question is available in an external corpus (Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022). A training instance $(x, y, D) \in X_{train}$ consists of a question x, an answer y, and the support set D, the documents that together support the answer. DAVID-GRPO automatically constructs these instances from a given document corpus (§4.1). A policy π_θ, a language model agent with parameters θ, answers x over several turns. In each turn it writes reasoning text and then either issues a search query, whose retrieved documents are appended to the context, or emits the final answer. We model this as an episodic MDP in the standard agentic RL setup (Appendix A). A step t is one turn, the action a_t is the query issued at step $t < T$ or the answer $\hat{y}$ at the terminal step T, and $\hat{D}_t$ denotes the documents retrieved at step t. A rollout produces a trajectory $\tau \sim \pi_\theta(x)$ that ends when the agent answers or exhausts a budget of tokens and tool calls, and training assigns it a reward $R(\tau; y, D)$. The policy parameters θ are updated with GRPO (Shao et al., 2024). For each query x, GRPO samples a group of G trajectories $\{\tau_1, \ldots, \tau_G\}$ from the policy before the update, $\pi_{\theta_{old}}$, standardizes their rewards within the group into advantages $\hat{A}_i = (R(\tau_i) - \text{mean}_j R(\tau_j))/\text{std}_j R(\tau_j)$, and updates π_θ by maximizing the clipped surrogate $f_{clip}(\rho_i(\theta), \hat{A}_i)$ with probability ratio $\rho_i(\theta)=\pi_\theta(\tau_i)/\pi_{\theta_{old}}(\tau_i)$, averaged over the group and regularized by a KL term $\beta\,\mathbb{D}_{KL}$ against a reference model (Eq. (6) in Appendix A).

3.2 EXPERT TRAJECTORY SEEDING

With few rollouts per question, standard GRPO rarely samples a rewarded trajectory and undergoes retrieval collapse (the no-seeding baseline in §5). Unlike SFT-based initialization (Zhang et al., 2025), which uses thousands of expert trajectories, expert trajectory seeding uses a handful of expert trajectories inside RL updates. Formally, the expert trajectory set $X_{\text{exp}}=\{(x_j, \tau_j^*) \mid \tau_j^* \sim \pi^*(x_j)\}_{j=1}^k$ holds k trajectories generated by an expert policy π^* (e.g., a larger model or a human). In our experiments, X_{exp} holds four HotpotQA trajectories released by AutoCoA (Appendix H), the only annotated examples that DAVID-GRPO uses and 99.9% fewer (4 vs. 5,276) than the SFT stage of AutoCoA. Appendix E.3 reports how performance changes with larger k. Seeding forms the early phase of training, in which every group is built from a question in X_{exp}. Training then continues on

X_{train} with on-policy groups only (50 and 215 steps in our experiments, Appendix E). As illustrated in Figure 2a, for an input x with $(x, \tau^*) \in X_{\text{exp}}$ we seed a group of size G with one off-policy expert trajectory τ^* and fill it with $G-1$ on-policy trajectories $\{\tau_2, \ldots, \tau_G\}$ sampled from $\pi_{\theta_{old}}$. To account for the distributional shift, we write the expert trajectory ratio as a product of the standard policy ratio and an additional expert-policy correction: $\rho^*_{\text{mix}}(\theta) = \omega^*(\theta) \cdot \frac{\pi_\theta(\tau^*)}{\pi_{\theta_{old}}(\tau^*)}$ where $\omega^*(\theta)$ accounts for the gap between the expert policy and the rollout policy. The expert trajectory receives the same reward $R(\tau^*; y, D)$ as the on-policy trajectories, and the advantage $\hat{A}(\tau^*)$ is computed by standardizing rewards within the mixed group, as in Appendix A. Applying the same clipped surrogate function f_{clip} and KL penalty $\mathbb{D}_{KL}$ as in Eq. 6, the mixed objective is:

$$\mathcal{L}_{Mixed}(\theta) = \frac{1}{G}\left(f_{clip}(\rho^*_{\text{mix}}(\theta), \hat{A}(\tau^*)) + \sum_{i=2}^{G} f_{clip}(\rho_i(\theta), \hat{A}(\tau_i)) \right) - \beta \mathbb{D}_{KL}. \tag{1}$$

The expert trajectory gives a gradient signal even when the agent's own exploration fails, and the on-policy trajectories let a high-reward trajectory of the agent itself receive a positive $\hat{A}_i$ beyond imitating τ^*. Figure 3 shows that this outperforms SFT on X_{exp} and pure off-policy RL. In implementation, we keep the standard policy ratio $\pi_\theta(\tau^*)/\pi_{\theta_{old}}(\tau^*)$ for τ^* but set the additional expert-policy correction $\omega^*(\theta)$ to 1. Appendix E.3 compares this choice with trajectory- and action-level corrections.

3.3 Evidence-Guided Continuation

Evidence-guided continuation scores each trajectory by evidence coverage (§3.3.1) and uses the score to resume the most promising partial trajectory when a group has no fully rewarded trajectory (§3.3.2).

3.3.1 Evidence-Coverage Reward

Expert trajectory seeding provides trajectories with non-zero reward early in training, but with the outcome reward alone small agents still fall into retrieval collapse (see Tree-GRPO in Table 3). A denser reward should score whether retrieval found the evidence. Aligning each retrieval step to a particular document would need step-level annotation, so we score the evidence that a trajectory has accumulated as a whole. To this end, we introduce an *evidence-coverage reward* (r_g). As illustrated in Figure 2b, let $\hat{\mathcal{D}}_{union}$ be the set of all documents retrieved in the trajectory, the union over intermediate steps ($t < T$):

$$\hat{\mathcal{D}}_{union} = \bigcup_{t=1}^{T-1} \hat{D}_t. \tag{2}$$

The evidence-coverage reward r_g is the recall of the support set D within this set:

$$r_g(\tau) = \frac{|\hat{\mathcal{D}}_{union} \cap D|}{|D|}. \tag{3}$$

Recall does not penalize an unsuccessful search while required evidence is still missing, and Appendix E.4 compares this choice with an F1-oriented reward.

We combine r_g with the outcome reward r_o (which evaluates the predicted answer $\hat{y}$ against y). The total reward is a weighted sum:

$$R(\tau) = \lambda r_g(\tau) + (1-\lambda) r_o(\tau, y) \tag{4}$$

where λ is a balancing hyperparameter. We set λ=0.5 in the main experiments and report sensitivity in Appendix E.4.

3.3.2 Prefix Continuation

With few rollouts, a group often contains no trajectory that reaches the maximum reward ($R(\tau) < 1$ for every τ in the group), so the advantages $\hat{A}_i$ carry little signal. Even after expert trajectory seeding, such groups still make up 20 to 33% of a batch during early training (Appendix B.8). We introduce

prefix continuation, which keeps the prefix of the group's highest-coverage trajectory up to the last retrieval that added evidence and resamples the rest of the trajectory on policy.

As illustrated in Figure 2b, within a sampled group of G trajectories, we identify the highest-reward trajectory τ_M and the lowest-reward trajectory τ_m. If even τ_M does not reach the maximum reward (i.e., $R(\tau_M) < 1$), we hypothesize that the remaining necessary information is not recalled, known as *bounded recall* (Covington et al., 2016) in IR. For such *near-miss* groups, we locate the last step t' at which coverage increased:

$$t' = \max\{t \mid r_g(\tau_{M,\leq t}) > r_g(\tau_{M,\leq t-1})\} \tag{5}$$

where $\tau_{M,\leq t}$ is a truncated trajectory of τ_M from step 1 to t, and r_g is the evidence-coverage reward. We then truncate τ_M at step t', keeping the history up to that retrieval, and resample the rest of the trajectory to obtain a new candidate $\tau'_M \sim \pi_\theta(x|\tau_{M,\leq t'})$.

If the continuation earns a higher reward than the original trajectory ($R(\tau'_M) > R(\tau_M)$), we replace τ_m with τ'_M, which keeps the group size fixed.

4 Experiments

4.1 Experimental Setup

Backbones. We train agents with at most 1.5B parameters, **Qwen2.5** (0.5B, 1.5B) (Qwen et al., 2025), **Llama-3.2-1B** (Grattafiori et al., 2024), and **Qwen3-0.6B** (Yang et al., 2025) (Appendix B.4), and additionally **Qwen2.5-7B** under the same budget to check whether the gains hold at a larger scale. We train the Qwen models from their base checkpoints, except for Tree-GRPO, which we train from the instruct versions because it degenerates from the base checkpoints (Appendix B.7). For Llama-3.2-1B we use the instruct version, because the base checkpoint ships without a chat template and cannot be trained with the agent prompt.

Training Data. DAVID-GRPO trains on 4,000 instances constructed automatically from the Wikipedia corpus released with HotpotQA (Yang et al., 2018). For the analysis in §5, we also train DAVID-GRPO on 5,276 HotpotQA training examples, with the support set D taken from the supporting documents annotated in HotpotQA. DAVID-GRPO needs the support set D of each training question, so we construct it from the corpus. We sample a document chain from the hyperlink graph, let Qwen2.5-3B write a multi-hop question backward along the chain, and keep the instance only if the chain is necessary and sufficient to answer the question (Appendix E.6). The sampled chain is the support set D, and 45% of the instances require three or more hops. Four expert trajectories from AutoCoA form the expert trajectory set (§3.2). Baselines use their released training configurations (Appendix E).

Training Budget. We train on 4×NVIDIA RTX 3090 GPUs (24GB VRAM) with a global batch size of 24 and 6 rollouts per example on average,[2] 144 rollouts per optimization step, and call this setting the *Low Training Budget*. The *High Training Budget* of Tree-GRPO uses 8×H20 GPUs (96GB VRAM) and 3,072 rollouts per step (512×6), so the low budget is 4.7% of the high budget. The reported 7B scores of StepSearch and Search-R1-v0.3 in Table 2 come from batch sizes of 256 and 512. Qwen2.5-7B does not fit on the RTX 3090s, so we train it on 8×NVIDIA RTX A6000 GPUs (48GB VRAM) with the same batch size and rollout budget as the smaller models. Appendix B.1 reports training cost and inference-time token counts, and Appendix E the remaining implementation details.

Baselines. We compare with three RL methods for search agents under the same low budget. **Tree-GRPO** (Ji et al., 2026) is the only method that reports 1.5B results under a high budget, so we also compare with the reported scores. **Search-R1-v0.3** (Jin et al., 2025a) trains with a *retrieval reward*, and **StepSearch** (Zheng et al., 2025) with step-wise lexical similarity rewards (§2). Appendix B.3 compares with **AutoCoA** (Zhang et al., 2025), which initializes from full SFT on a large trajectory set, and Appendix B.11 reports no-training controls of the base model.

[2] DAVID-GRPO samples five full rollouts per example, and prefix continuation adds 1.1 rollouts per example on average that resume from a prefix of an existing trajectory (Appendix E).

Table 1: EM and F1 on six multi-hop QA benchmarks. Best results are **bolded** and second-best underlined. **Red** marks low-budget EM above the best high-budget baseline. † denotes results from Ji et al. (2026). DAVID-GRPO rows are brown. *Search* is the average number of search actions per question over the six benchmarks. ‡ is the final number of valid searches per rollout in the released training logs of Ji et al. (2026) (0.998 for Tree-GRPO, 0.0 for GRPO). Teal Antileak-m cells are contamination-free because the questions post-date the knowledge cutoff of the backbone. Table 5 reports DAVID-GRPO trained with annotated evidence.

	Method	HotpotQA		2wiki		Musique		Bamboogle		Bamtwoogle		Antileak-m		Avg.		
		EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	Search
	NO TRAINING															
	Direct Inference†	5.9	-	4.3	-	2.6	-	8.0	-	-	-	-	-	-	-	-
	HIGH TRAINING BUDGET: up to batch size 512×6 rollout per instance $= 3,072$															
Qwen2.5-1.5B	Search-o1†	11.6	-	12.2	-	3.1	-	13.0	-	-	-	-	-	-	-	-
	GRPO†	14.6	-	24.4	-	2.2	-	4.0	-	-	-	-	-	-	-	0.00‡
	Tree-GRPO†	29.5	-	26.8	-	6.6	-	13.6	-	-	-	-	-	-	-	1.00‡
	LOW TRAINING BUDGET: up to batch size 24×6 rollout per instance $= 144$															
	Tree-GRPO	12.9	18.6	20.5	23.3	2.1	7.2	12.0	15.5	5.0	8.2	11.7	15.7	10.7	14.8	0.00
	StepSearch	11.9	18.1	13.5	18.1	2.2	6.9	3.2	8.7	4.0	6.0	12.1	17.3	7.8	12.5	1.11
	Search-R1-v0.3 *w/ retrieval reward*	19.0	26.5	21.3	26.4	3.6	9.2	8.0	14.3	3.0	5.5	16.7	22.2	11.9	17.4	0.45
	DAVID-GRPO	**23.5**	**32.2**	**26.7**	**32.0**	**8.0**	**14.0**	**18.4**	**28.1**	**22.0**	**26.2**	**36.9**	**41.4**	**22.6**	**29.0**	**3.13**
Llama-3.2-1B	Tree-GRPO	12.4	18.4	**20.5**	**23.4**	1.6	7.1	4.0	8.9	**9.0**	**11.9**	16.0	20.7	10.6	15.1	0.00
	StepSearch	16.0	**24.5**	11.7	16.2	2.7	**9.7**	4.0	11.0	3.0	5.7	22.4	**31.4**	10.0	16.4	1.14
	Search-R1-v0.3 *w/ retrieval reward*	8.5	12.4	13.5	16.2	0.7	3.7	0.8	1.6	2.0	4.7	13.0	15.3	6.4	9.0	0.09
	DAVID-GRPO	**16.7**	24.0	12.6	17.5	**2.9**	6.9	**8.8**	**16.0**	3.0	6.3	**24.4**	31.3	**11.4**	**17.0**	**1.77**
Qwen2.5-0.5B	Tree-GRPO	9.2	12.8	**20.9**	22.8	0.7	3.8	3.2	5.2	4.0	5.5	10.3	12.7	8.1	10.5	0.00
	StepSearch	2.1	4.5	4.3	7.3	0.2	1.8	1.6	2.1	0.0	0.6	0.2	0.8	1.4	2.9	0.89
	Search-R1-v0.3 *w/ retrieval reward*	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00
	DAVID-GRPO	**13.1**	**19.2**	20.5	**24.5**	**2.3**	**6.0**	**4.0**	**8.0**	**5.0**	**6.3**	**16.0**	**22.7**	**10.2**	**14.4**	**2.60**

Benchmarks and Metrics. We evaluate on six multi-hop QA benchmarks. **HotpotQA** (Yang et al., 2018) covers *bridge* and *comparison* questions with two hops. **2WikiMultiHopQA** (Ho et al., 2020) and **MuSiQue** (Trivedi et al., 2022) add compositional questions, up to four hops in MuSiQue. **Bamboogle** (Press et al., 2023) and **BamTwoogle** (Aksitov et al., 2024) are handcrafted benchmarks with 2-hop and 2- to 4-hop questions. **AntiLeakBench-Multihop** (Wu et al., 2025) (2024 subset, 2 to 3 hops), which we call Antileak-m, contains questions that post-date the knowledge cutoff of the backbones and measures contamination-free performance (Appendices C and D). We report **Exact Match (EM)** and **F1**, and Appendix B.10 reports LLM-as-judge evaluation. Appendix B.6 additionally evaluates DAVID-GRPO on the multi-turn long-context subset of BFCL, a tool-use benchmark outside Wikipedia QA.

4.2 EXPERIMENTAL RESULTS

Matching High-Budget Training. Table 1 first compares DAVID-GRPO with the results reported for Qwen2.5-1.5B under the high training budget of Tree-GRPO, 3,072 rollouts per step on 8×H20 GPUs. With **4.7%** of those rollouts on four RTX 3090s, DAVID-GRPO averages **19.2 EM against 19.1** for high-budget Tree-GRPO on the four benchmarks it reports, exceeding it on MuSiQue and Bamboogle (red), trailing on HotpotQA (23.5 vs. 29.5), and tying on 2WikiMultiHopQA (26.7 vs. 26.8). High-budget vanilla GRPO and Search-o1 average 11.3 and 10.0 EM on the same benchmarks, so the low-budget DAVID-GRPO also exceeds them by 8 to 9 points (Figure 1).

Low-Budget Comparison. Under the low budget, DAVID-GRPO reaches **22.6 average EM and 29.0 F1** with Qwen2.5-1.5B against 11.9 and 17.4 for the best baseline, Search-R1-v0.3, and leads on every benchmark. The margin is largest on the contamination-free Antileak-m, 20.2 EM points. Moving from 3,072 to 144 rollouts costs Tree-GRPO EM on all four benchmarks it reports, and GRPO without seeding earns no reward at all (§5.2). DAVID-GRPO also has the highest average EM on Qwen2.5-0.5B and Llama-3.2-1B, where Search-R1-v0.3 collapses to 0 EM at 0.5B. On the

Table 2: Scaling to Qwen2.5-7B. EM and F1 with the Qwen2.5-7B backbone. High-budget rows are reported scores from the original papers. † Ji et al. (2026) uses batch size 512, ‡ Zheng et al. (2025) Qwen2.5-7B-Base with batch size 256 on 16×H800, and § Jin et al. (2025a) Qwen2.5-7B-Base GRPO with retrieval reward and batch size 512 on 8×H100. Best and second-best results among low-budget methods are **bolded** and underlined. *Search* is the average number of search actions per question over the six benchmarks. ¶ is measured on HotpotQA with the released best checkpoint.

Method	HotpotQA		2wiki		Musique		Bamboogle		Bamtwoogle		Antileak-m		Avg.		
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	Search
NO TRAINING															
Direct Inference†	18.3	-	25.0	-	3.1	-	12.0	-	-	-	-	-	-	-	-
HIGH TRAINING BUDGET (reported)															
Search-o1†	22.1	-	21.8	-	5.4	-	32.0	-	-	-	-	-	-	-	-
GRPO†	42.5	-	40.7	-	19.1	-	43.2	-	-	-	-	-	-	-	-
Tree-GRPO†	44.6	-	42.3	-	20.2	-	44.0	-	-	-	-	-	-	-	1.82¶
StepSearch‡	38.0	49.3	38.5	45.0	21.6	32.4	46.7	57.3	-	-	-	-	-	-	-
Search-R1-v0.3 *w/ retrieval reward*§	41.6	-	37.5	-	16.4	-	38.7	-	-	-	-	-	-	-	-
LOW TRAINING BUDGET: up to batch size 24×6 rollout per instance $= 144$															
Tree-GRPO	**35.5**	**46.4**	28.8	35.0	**15.3**	**23.5**	40.0	51.1	38.0	**46.5**	43.5	52.6	33.5	42.5	1.92
StepSearch	27.5	37.7	18.2	25.4	12.9	21.5	31.2	44.0	36.0	40.7	43.3	53.0	28.2	37.1	2.40
Search-R1-v0.3 *w/ retrieval reward*	34.9	45.5	35.5	41.3	13.0	21.5	34.4	45.7	36.0	42.5	42.4	47.1	32.7	40.6	1.73
DAVID-GRPO	35.2	45.7	**38.7**	**45.8**	**15.3**	23.2	**43.2**	**52.6**	**42.0**	45.6	**49.2**	**54.7**	**37.3**	**44.6**	**3.15**

more recent Qwen3-0.6B, StepSearch ties DAVID-GRPO at 19 average EM while Tree-GRPO and Search-R1-v0.3 stay below 10 (Appendix B.4).

Multi-Turn Retrieval Under the Low Budget. The Search column of Table 1 shows that DAVID-GRPO issues the most searches per question on every backbone. With Qwen2.5-1.5B, DAVID-GRPO averages 3.1 searches over the six benchmarks, whereas Tree-GRPO, Search-R1-v0.3, and StepSearch average 0.0, 0.5, and 1.1. On Llama-3.2-1B and Qwen2.5-0.5B, DAVID-GRPO averages 1.8 and 2.6 searches, and no baseline exceeds 1.1. Even 3,072 rollouts per step do not keep an outcome-reward agent retrieving. The released training log of high-budget GRPO ends with 0.0 valid searches per rollout, and that of high-budget Tree-GRPO with 1.0.

Scaling to a 7B Agent. With Qwen2.5-7B under the same budget (Table 2), DAVID-GRPO has the highest average EM and F1 (37.3 and 44.6 vs. 33.5 and 42.5 for Tree-GRPO), the largest margin on Antileak-m (49.2 vs. 43.5), and the most searches per question (3.2 vs. 1.7 to 2.4 over the six benchmarks). Even the high-budget Tree-GRPO checkpoint searches less. On HotpotQA, the released 7B checkpoint averages 1.8 searches per question, and DAVID-GRPO averages 3.0 right after expert trajectory seeding, before the main GRPO stage (Appendix B.7). Against the reported high-budget scores, DAVID-GRPO reaches 88% of Tree-GRPO and 91% of StepSearch on the four shared benchmarks with 4.7% of the rollouts.

5 ANALYSIS

This section separates the two sources of the gains. In §5.1 and §5.2, DAVID-GRPO is trained on the HotpotQA training set with its annotated evidence (gold), the same kind of supervision the retrieval-reward baselines use, so the comparisons isolate the RL mechanisms. §5.3 then keeps the mechanisms fixed and changes only the training data.

5.1 MULTI-HOP RETRIEVAL CAPABILITIES

Retrieval Behavior. Table 3 compares how the agents search on HotpotQA and MuSiQue. DAVID-GRPO issues about three searches per question and reaches the highest document recall, both for any support document (*Any*) and for the complete support set (*All*). Tree-GRPO stops searching entirely and Search-R1-v0.3 averages half a search per question, so both remain in retrieval collapse, and StepSearch searches once. The additional searches of DAVID-GRPO mostly target missing evidence.

Table 3: Retrieval behavior. Retrieval statistics and EM (%) on HotpotQA and MuSiQue. *Search* is the average number of search actions per question. *Prec.* is the percentage of retrieved documents that belong to the annotated support set D, *Any* is the percentage of the documents of D that were retrieved, pooled over questions, and *All* is the percentage of questions for which every document of D was retrieved. *Unnec.* is the percentage of searches issued after the whole support set has been retrieved. The backbone model is Qwen2.5-1.5B except for Tree-GRPO with Qwen2.5-1.5B-Inst.

Method	HotpotQA						MuSiQue					
	Search	Prec.	Any	All	Unnec.	EM	Search	Prec.	Any	All	Unnec.	EM
Tree-GRPO	0.00	-	0.0	0.0	-	12.9	0.00	-	0.0	0.0	-	2.1
StepSearch	1.12	39.6	40.3	13.8	1.6	11.9	1.11	17.2	14.7	1.8	0.2	2.2
Search-R1-v0.3	0.53	**40.1**	20.9	6.7	0.0	19.0	0.59	**18.3**	8.7	1.3	0.0	3.6
DAVID-GRPO	3.05	28.2	**52.3**	**31.6**	9.8	**24.8**	3.48	14.3	**25.9**	**8.2**	2.6	**7.1**

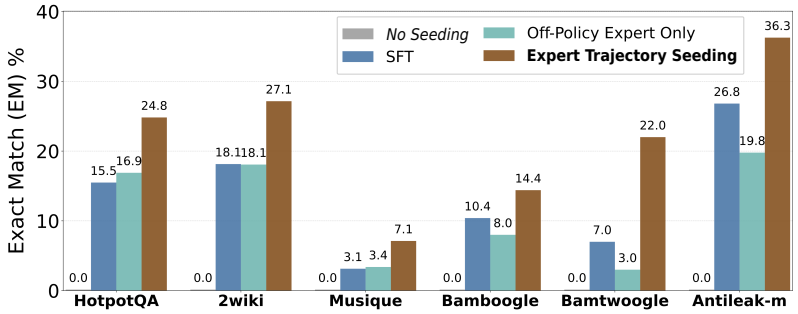


Figure 3: Analysis on expert trajectory seeding. Performance comparison of different ways to use the four expert trajectories of X_{exp} followed by GRPO with the *evidence-coverage reward* on Qwen2.5-1.5B.

Only 9.8% on HotpotQA and 2.6% on MuSiQue are issued after the support set is already complete, and document precision stays within 12 points of the best baseline. Appendix B.12 breaks recall down into bridge and answer documents, and Appendix B.2 reports HotpotQA question types and MuSiQue hop depths.

5.2 Ablation Study

Impact of Expert Trajectory Seeding. Figure 3 compares different ways to use the same four expert trajectories. Without them, the agent undergoes retrieval collapse and scores zero EM on every benchmark. Expert trajectory seeding, which combines off-policy trajectories from X_{exp} (§3.2) with on-policy generations, achieves the highest performance. When using X_{exp} only, standard SFT performs comparably to or better than off-policy RL. Adding on-policy rollouts gives a larger gain than either SFT or off-policy RL alone, which supports expert trajectory seeding. Keeping the same four trajectories but replacing the rest of DAVID-GRPO with vanilla GRPO, StepSearch, Search-R1-v0.3, or Tree-GRPO gives at most 20.4 average EM against 22.0 (Appendix E.3), so the trajectories alone do not explain the gain.

Effectiveness of Evidence-Guided Continuation. Table 4 isolates the evidence-guided continuation components. The full DAVID-GRPO system gives the best results across all six benchmarks. Removing *prefix continuation* reduces performance, and removing both *prefix continuation* and *evidence-coverage reward* gives the lowest results. The variant without both is trained with the outcome reward alone, so the lowest results on all six benchmarks indicate that the outcome reward is insufficient under the low budget. Appendix B.12 shows that the two components raise document recall cumulatively (Table 22).

Appendix E.4 compares our evidence-coverage reward with StepSearch and Search-R1-v0.3 rewards, which do not explicitly reward coverage of the support set.

Table 4: Ablation on evidence-guided continuation. Performance of DAVID-GRPO and its ablations on Qwen2.5-1.5B across multi-hop QA benchmarks. The *w/o coverage reward* variant removes both components and is trained with only the outcome reward.

Method	HotpotQA		2wiki		Musique		Bamboogle		Bamtwoogle		Antileak-m		Avg.	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
DAVID-GRPO	**24.8**	**33.8**	**27.2**	**32.3**	**7.1**	**12.6**	**14.4**	**24.2**	**22.0**	**25.4**	**36.3**	**41.1**	**22.0**	**28.2**
w/o continuation	23.3	31.9	24.4	30.0	5.8	11.3	**14.4**	20.9	19.0	23.8	33.2	38.2	20.0	26.0
w/o coverage reward	20.8	28.9	22.0	27.0	4.5	10.1	10.4	19.9	10.0	12.7	31.4	36.0	16.5	22.4

Table 5: Gold versus constructed evidence. EM and F1 of DAVID-GRPO trained on the 5,276 HotpotQA training examples with annotated evidence (*gold*) and on the 4,000 instances constructed in §4.1 (*synth*, the default DAVID-GRPO). The higher value of each pair is **bolded**. *Search* is the average number of search actions per question over the six benchmarks. Teal Antileak-m cells are contamination-free evaluations.

Backbone	Evidence	HotpotQA		2wiki		Musique		Bamboogle		Bamtwoogle		Antileak-m		Avg.		
		EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	Search
Qwen2.5-7B	Gold	**35.2**	45.6	37.7	44.5	14.2	21.6	40.0	52.3	**44.0**	**47.1**	**55.2**	**59.8**	**37.7**	**45.1**	**3.95**
	Synth	**35.2**	**45.7**	**38.7**	**45.8**	**15.3**	**23.2**	**43.2**	**52.6**	42.0	45.6	49.2	54.7	37.3	44.6	3.15
Qwen2.5-1.5B	Gold	**24.8**	**33.8**	**27.2**	**32.3**	7.1	12.6	14.4	24.2	**22.0**	25.4	36.3	41.1	22.0	28.2	3.07
	Synth	23.5	32.2	26.7	32.0	**8.0**	**14.0**	**18.4**	**28.1**	**22.0**	**26.2**	**36.9**	**41.4**	**22.6**	**29.0**	**3.13**
Llama-3.2-1B	Gold	**17.7**	**25.2**	**16.1**	**21.4**	**3.2**	**8.5**	8.0	14.5	**3.0**	**6.6**	23.3	**31.7**	**11.9**	**18.0**	1.30
	Synth	16.7	24.0	12.6	17.5	2.9	6.9	**8.8**	**16.0**	**3.0**	6.3	**24.4**	31.3	11.4	17.0	**1.77**
Qwen2.5-0.5B	Gold	10.8	16.0	17.4	20.8	2.0	5.3	**4.8**	**8.1**	**6.0**	**7.0**	10.6	14.4	8.6	11.9	**2.64**
	Synth	**13.1**	**19.2**	**20.5**	**24.5**	**2.3**	**6.0**	4.0	8.0	5.0	6.3	**16.0**	**22.7**	**10.2**	**14.4**	2.60

5.3 GOLD VERSUS CONSTRUCTED EVIDENCE

Table 5 compares DAVID-GRPO trained on the HotpotQA training set with annotated evidence (gold) and on the constructed instances (synth). Average EM differs by at most 1.6 points on any backbone, and the direction of the difference depends on the backbone. The learned retrieval behavior is also similar. The two variants differ by at most 0.5 searches per question on every backbone except Qwen2.5-7B, where gold issues 0.8 more searches per question. Per benchmark, synth trails gold by about one EM point on HotpotQA, the training distribution of gold, on Qwen2.5-1.5B and Llama-3.2-1B, and leads on Bamboogle and MuSiQue on the two largest backbones.

6 CONCLUSION

We studied RL for multi-hop retrieval agents under the compute budget of commodity GPUs, where each update sees only a few rollouts per question. Under this low budget most sampled trajectories retrieve none of the required evidence, and RL from the outcome reward alone leaves a small agent in retrieval collapse, answering without retrieval. DAVID-GRPO increases the learning signal obtained from each update in two ways. Expert trajectory seeding places a handful of expert trajectories in each GRPO group so that rewarded trajectories exist from the first update, and evidence-guided continuation rewards evidence coverage and spends additional rollouts on resuming the most promising partial trajectory, with the evidence automatically constructed from the corpus graph. On six multi-hop QA benchmarks, DAVID-GRPO trained with 144 rollouts per step on four RTX 3090 GPUs nearly doubles the average EM of the best low-budget baseline at 1.5B parameters, matches Tree-GRPO trained with 20 times more rollouts, and carries the gains to backbones from 0.5B to 7B parameters. Under the low budget, DAVID-GRPO teaches the agent multi-turn retrieval where the baselines do not. On six benchmarks whose questions require at least two hops, Qwen2.5-1.5B trained with DAVID-GRPO issues 3.1 searches per question, whereas Tree-GRPO, Search-R1-v0.3, and StepSearch issue 0.0, 0.5, and 1.1.

Ethics Statement

This work studies how to train retrieval and tool-using agents with smaller models and commodity GPUs. Lowering the compute barrier can make agent research easier to reproduce and extend outside large industrial labs. The same techniques could also improve agents used for harmful information gathering or automated misuse if deployed without safeguards. The constructed training data is derived from Wikipedia articles and contains no personal or private information beyond the source articles. Deployment of such agents requires the usual access control, monitoring, and tool-use safeguards for agentic systems.

Reproducibility Statement

Hyperparameters, hardware, and the training and evaluation pipelines are listed in Appendix E. The corpus-graph evidence construction pipeline is described in §4.1, with the hop distribution and filtering details in Appendix E.6 and the generation prompts in Appendix F.1. The four expert trajectories used for bootstrapping are given in Appendix H, and the agent and judge prompts in Appendix F. Benchmark statistics and knowledge cutoff dates are in Appendices C and D. Training and evaluation code is available at the anonymous repository linked in the abstract.

AI Use Statement

Language models are part of the method and the evaluation. Qwen2.5-3B generates the training questions and referring expressions of the corpus-graph evidence construction pipeline (§4.1), and Claude Sonnet 5 serves as the judge in the LLM-as-judge evaluation (Appendix B.10). We also used a language model to edit the manuscript for readability and to format tables. All experiments, analyses, and claims were designed, run, and verified by the authors.

References

Renat Aksitov, Sobhan Miryoosefi, Zonglin Li, Daliang Li, Sheila Babayan, Kavya Kopparapu, Zachary Fisher, Ruiqi Guo, Sushant Prakash, Pranesh Srinivasan, Manzil Zaheer, Felix Yu, and Sanjiv Kumar. ReST meets react: Self-improvement for multi-step reasoning LLM agent. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024. URL https://openreview.net/forum?id=7xknRLr7QE.

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. Synthetic QA corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6168–6173, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1620. URL https://aclanthology.org/P19-1620.

Tamay Besiroglu, Sage Andrus Bergerson, Amelia Michael, Lennart Heim, Xueyun Luo, and Neil Thompson. The compute divide in machine learning: A threat to academic contribution and scrutiny? *arXiv preprint arXiv:2401.02452*, 2024.

Yiqun Chen, Lingyong Yan, Weiwei Sun, Xinyu Ma, Yi Zhang, Shuaiqiang Wang, Dawei Yin, Yiming Yang, and Jiaxin Mao. Improving retrieval-augmented generation through multi-agent reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=9Ia0KiVAut.

Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, pp. 191–198, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340359. doi: 10.1145/2959100.2959190. URL https://doi.org/10.1145/2959100.2959190.

Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. Hierarchical graph network for multi-hop question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8823–8838, Online, November

2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.710. URL https://aclanthology.org/2020.emnlp-main.710.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580.

Yuelyu Ji, Rui Meng, Zhuochun Li, and Daqing He. Curriculum guided reinforcement learning for efficient multi hop retrieval augmented generation. *arXiv preprint arXiv:2505.17391*, 2025.

Yuxiang Ji, Ziyu Ma, Yong Wang, Guanhua Chen, Xiangxiang Chu, and Liaoni Wu. Tree search for llm agent reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2026.

Bowen Jin, Jinsung Yoon, Priyanka Kargupta, Sercan O. Arik, and Jiawei Han. An empirical study on reinforcement learning for reasoning-search interleaved LLM agents. In *The First Structured Knowledge for Large Language Models Workshop*, 2025a. URL https://openreview.net/forum?id=IQNZIBspz5.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan O Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training LLMs to reason and leverage search engines with reinforcement learning. In *Second Conference on Language Modeling*, 2025b. URL https://openreview.net/forum?id=Rwhi91ideu.

Jiajie Jin, Yutao Zhu, Zhicheng Dou, Guanting Dong, Xinyu Yang, Chenghao Zhang, Tong Zhao, Zhao Yang, and Ji-Rong Wen. Flashrag: A modular toolkit for efficient retrieval-augmented generation research. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, pp. 737–740, New York, NY, USA, 2025c. Association for Computing Machinery. ISBN 9798400713316. doi: 10.1145/3701716.3715313. URL https://doi.org/10.1145/3701716.3715313.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

Helena Kudiabor. Academics lack access to powerful chips needed for AI research. *Nature*, 2024.

Shishir G Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The berkeley function calling leaderboard (BFCL): From tool use to agentic evaluation of large language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=2GmDdhBdDk.

Akshara Prabhakar, Zuxin Liu, Ming Zhu, Jianguo Zhang, Tulika Manoj Awalgaonkar, Shiyu Wang, Zhiwei Liu, Haolin Chen, Thai Quoc Hoang, Juan Carlos Niebles, Shelby Heinecke, Weiran Yao, Huan Wang, Silvio Savarese, and Caiming Xiong. APIGen-MT: Agentic pipeline for multi-turn data generation via simulated agent-human interplay. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=qk6ORqQ4Cu.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=feiAVaSXdb.

Qwen, :, 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, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. R1-searcher: Incentivizing the search capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2503.05592*, 2025. URL https://arxiv.org/abs/2503.05592.

Zhongxiang Sun, Qipeng Wang, Weijie Yu, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Yang Song, and Han Li. ReARTeR: Retrieval-augmented reasoning with trustworthy process rewarding. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2025. URL https://doi.org/10.1145/3726302.3730102.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

Xiaobao Wu, Liangming Pan, Yuxi Xie, Ruiwen Zhou, Shuai Zhao, Yubo Ma, Mingzhe Du, Rui Mao, Anh Tuan Luu, and William Yang Wang. AntiLeakBench: Preventing data contamination by automatically constructing benchmarks with updated real-world knowledge. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18403–18419, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.901. URL https://aclanthology.org/2025.acl-long.901/.

Wenhan Xiong, Mo Yu, Xiaoxiao Guo, Hong Wang, Shiyu Chang, Murray Campbell, and William Yang Wang. Simple yet effective bridge reasoning for open-domain multi-hop question answering. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pp. 48–52, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5806. URL https://aclanthology.org/D19-5806.

Wenhan Xiong, Xiang Li, Srini Iyer, Jingfei Du, Patrick Lewis, William Yang Wang, Yashar Mehdad, Scott Yih, Sebastian Riedel, Douwe Kiela, and Barlas Oguz. Answering complex open-domain questions with multi-hop dense retrieval. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=EMHoBG0avc1.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2369–2380, 2018.

Yuxiang Zhang, Yuqi Yang, Jiangming Shu, Xinyan Wen, and Jitao Sang. Agent models: Internalizing chain-of-action generation into reasoning models. *arXiv preprint arXiv:2503.06580*, 2025.

Xuhui Zheng, Kang An, Ziliang Wang, Yuhang Wang, and Yichao Wu. StepSearch: Igniting LLMs search ability via step-wise proximal policy optimization. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21816–21841, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1106. URL https://aclanthology.org/2025.emnlp-main.1106/.

A FULL MDP FORMULATION

This section provides the full specification of the MDP summarized in §3.1. We conceptualize the reasoning process of π_θ as a Markov decision process (MDP) $\langle S, A, P, R \rangle$,[3] where S and A are sets of states and actions, respectively. Let $a \oplus b$ signify the concatenation of strings a and b. The state transition is formulated as $s_{t+1} = s_t \oplus a_t \oplus \hat{D}_t$, where a_t for $t < T$ is an intermediate query for retrieval, $\hat{D}_t$ is the set of retrieved documents queried by a_t,[4] and the terminal action emits the predicted answer $\hat{y}$.[5] The environment dynamics are deterministic, where the next state is formed by concatenating the chosen action to the current sequence, i.e., $P(s_{t+1}|s_t, a_t) = 1$. For each rollout during on-policy RL and inference time, π_θ produces a trajectory $\tau = s_0 \oplus \cdots \oplus s_T$ given by a user question x and the task prompt s_0, i.e., $\tau \sim \pi_\theta(x)$. During training, given a training instance $(x, y, D) \in X_{train}$, a reward $R(\tau; y, D)$ is calculated by the reward function R where y and D are respectively the answer and the evidence to be retrieved. The policy parameters θ are then updated with the GRPO objective. Defining the probability ratio $\rho_i(\theta) = \frac{\pi_\theta(\tau_i)}{\pi_{\theta_{old}}(\tau_i)}$, the clipped surrogate function $f_{clip}(\rho, A) = \min(\rho A, \text{clip}(\rho, 1-\epsilon, 1+\epsilon)A)$ with a clipping parameter ϵ, and the KL divergence term against the reference model π_{ref} as $\mathbb{D}_{KL}$, the objective is:

$$\mathcal{L}_{GRPO}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G} f_{clip}(\rho_i(\theta), \hat{A}_i)\right] - \beta\mathbb{D}_{KL}. \tag{6}$$

B FURTHER ANALYSIS

Unless noted otherwise, the analyses in this appendix use the gold variant of DAVID-GRPO, as in §5.

B.1 ADDITIONAL COMPUTATION ANALYSIS

Table 6 reports the training cost of DAVID-GRPO and its ablations in GPU-days measured on RTX 3090 GPUs. Under a fixed peak memory budget of 20,890 MiB per GPU and no CPU offload, DAVID-GRPO takes 9.96 GPU-days, compared with 8.07 without prefix continuation and 8.74 without evidence-coverage reward. On four RTX 3090 GPUs, this adds less than 0.5 wall-clock days over the ablations.

Table 6: Training cost measured on RTX 3090 GPUs with a fixed peak memory budget of 20,890 MiB per GPU and no CPU offload.

Method	GPU-days (RTX 3090)
DAVID-GRPO	**9.96**
w/o continuation	8.07
w/o coverage reward	8.74

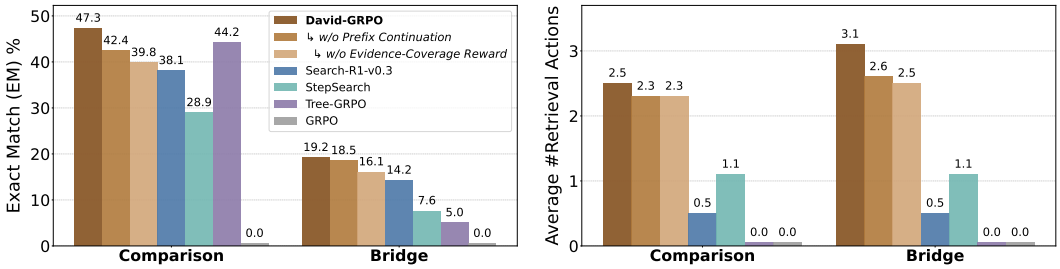
Table 7: Average number of retrieved tokens at inference time with Qwen2.5-1.5B.

Method	HotpotQA	2wiki	Musique	Bamboogle	Bamtwoogle	Antileak-m	Avg.
DAVID-GRPO	**1468.1**	**1595.6**	**1647.2**	**1466.4**	**1537.1**	**1160.8**	**1479.2**
w/o continuation	1253.1	1443.4	1441.4	1229.6	1406.1	1074.8	1308.1
w/o coverage reward	1212.5	1456.0	1353.2	1108.9	1418.8	1139.6	1281.5
Search-R1-v0.3	245.6	177.6	270.0	323.9	234.2	267.6	253.2
StepSearch	550.6	591.1	529.3	534.1	554.7	468.7	538.1
Tree-GRPO	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Tables 7 and 8 report inference-time retrieved and output token counts. DAVID-GRPO uses more retrieved tokens than the baselines because it performs more retrieval actions. The lower token counts of the baselines reflect their retrieval behavior. Tree-GRPO skips retrieval, and the other baselines stop after shallow retrieval (Figures 4b and 5b).

Table 8: Average number of output tokens generated by Qwen2.5-1.5B at inference time.

Method	HotpotQA	2wiki	Musique	Bamboogle	Bamtwoogle	Antileak-m	Avg.
DAVID-GRPO	**479.7**	449.0	**537.9**	**455.2**	509.9	383.4	469.2
w/o continuation	454.0	433.4	532.4	440.7	511.6	368.6	456.8
w/o coverage reward	462.3	**465.9**	527.6	433.1	**530.9**	**441.8**	**476.9**
Search-R1-v0.3	325.2	412.2	307.2	223.6	371.4	256.6	316.0
StepSearch	126.7	126.0	127.1	109.7	174.6	117.9	130.3
Tree-GRPO	94.5	87.7	105.9	85.2	98.3	73.3	90.8



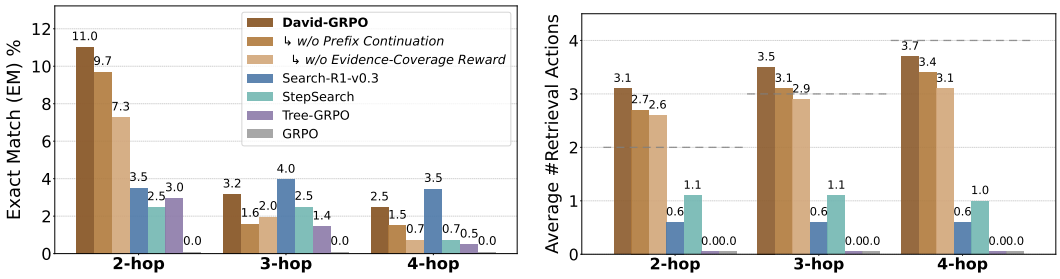
(a) EM by user question types.

(b) Avg. retrieval actions by user question types.

Figure 4: EM and average number of unique retrieval actions on HotpotQA by user question types with Qwen2.5-1.5B. Search-R1-v0.3 is trained along with its retrieval reward.

B.2 HotpotQA Question Types and MuSiQue Reasoning Depth

Reasoning Types in HotpotQA. We analyze HotpotQA performance across two reasoning types: *comparison* (comparing mentioned entities) and *bridge* (identifying missing intermediate entities). As shown in Figure 4a, DAVID-GRPO leads in both categories. Tree-GRPO performs well on *comparison* tasks but drops to the second-lowest on *bridge* tasks, where retrieval is needed to locate hidden entities. Figure 4b shows the retrieval behavior behind this gap. DAVID-GRPO and its ablations perform 2–3 retrieval steps. StepSearch averages 1.1 actions, and Tree-GRPO records zero actions, relying only on parametric knowledge (see Appendix G.2).



(a) EM by reasoning hops.

(b) Avg. retrieval actions by reasoning hops.

Figure 5: EM and average number of unique retrieval actions on MuSiQue by reasoning hops with Qwen2.5-1.5B. Dashed lines indicate the minimum number of retrieval actions required for each hop subset. Search-R1-v0.3 is trained along with its retrieval reward.

We stratify MuSiQue performance by reasoning hops. As shown in Figure 5a, DAVID-GRPO ranks first or second across all depths. Search-R1-v0.3 scores slightly higher on the 3- and 4-hop subsets, but Figure 5b shows that it averages only one retrieval action even for complex queries

[3] In this setting, the discount factor is set to 1, as the tasks are episodic and we are primarily concerned with the final correctness of the terminal state.

[4] $\hat{D}_t$ is converted into the textual sequence of documents when concatenated.

[5] $\hat{D}_T = \varnothing$ as no retrieval is conducted in step T.

(see Appendix G.4). This indicates that its higher-hop scores are not tied to multi-turn retrieval. In contrast, DAVID-GRPO increases retrieval depth with the annotated hop count.

Table 9: Controlled interventions on MuSiQue 3-hop and 4-hop subsets. EM and average number of retrieval actions with Qwen2.5-1.5B.

Method	EM		Retrieval	
	3-hop	4-hop	3-hop	4-hop
Search-R1-v0.3	4.0	3.5	0.6	0.6
DAVID-GRPO	3.2	2.5	3.5	3.7
+ *oracle answer doc append*	3.2	1.7	3.9	4.2
+ *oracle answer doc replace*	8.2	5.7	3.9	4.3
+ continuation only	3.3	1.0	**4.2**	4.3
+ early-hop truncation + continuation	**4.1**	**4.2**	4.1	**4.8**

We further analyze the 3-hop and 4-hop subsets in Table 9. The average retrieval depth of DAVID-GRPO increases only from 3.5 to 3.7 when moving from 3-hop to 4-hop questions, leaving little margin for missed evidence. Appending an oracle answer document does not improve EM, whereas replacing one final-hop retrieved document with it gives larger gains. Recall of later hops is not the only bottleneck. The gain from replacement points to the retrieved context as a second bottleneck, either through length or through distracting documents. Early-hop truncation followed by continuation improves 3-hop EM from 3.2 to 4.1 and 4-hop EM from 2.5 to 4.2.

B.3 FULL SFT WITH LARGE ANNOTATION SETS

Table 10: Comparison with Full SFT Initialization. Results across multi-hop QA benchmarks including EM and F1 scores. Results marked with † are reported from Ji et al. (2026). Scores marked in **red** indicate instances where methods under LOW TRAINING BUDGET outperform the best HIGH TRAINING BUDGET baseline. Base model is Qwen2.5-1.5B.

Method	HotpotQA		2wiki		Musique		Bamboogle		Bamtwoogle		Antileak-m		Avg.	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
HIGH TRAINING BUDGET														
Tree-GRPO†	29.5	-	26.8	-	6.6	-	13.6	-	-	-	-	-	-	-
LOW TRAINING BUDGET														
Full SFT	23.8	32.4	20.6	26.6	7.1	12.9	22.4	32.4	20.0	21.7	37.1	46.1	21.8	28.7
+ GRPO (AutoCoA)	32.1	42.6	27.2	34.5	11.2	17.6	32.8	**40.5**	23.0	29.4	45.9	52.1	28.7	36.1
+ DAVID-GRPO	**32.6**	**42.7**	**28.1**	**36.2**	**12.2**	**19.0**	**33.6**	39.4	**32.0**	**37.8**	**46.2**	**53.6**	**30.8**	**38.1**

We also test DAVID-GRPO with Full SFT initialization when large trajectory annotation sets are available. Table 10 shows that GRPO after Full SFT improves over the Full SFT baseline for both AutoCoA and DAVID-GRPO. Both AutoCoA and DAVID-GRPO surpass the high-budget Tree-GRPO baseline across the reported benchmarks (marked in red). DAVID-GRPO also outperforms AutoCoA on most benchmarks, with the largest gap on BamTwoogle: 9.0 pp in EM and 8.4 pp in F1.

B.4 EVALUATION ON QWEN3-0.6B

Table 11 repeats the low-budget comparison with Qwen3-0.6B (Yang et al., 2025), a backbone released in 2025. StepSearch ties DAVID-GRPO at 19 average EM, and both learn multi-turn retrieval, with 2.0 and 2.6 searches per question. Tree-GRPO and Search-R1-v0.3 stop searching on this backbone as well and stay below 10 EM. StepSearch sustains retrieval on only two of the five backbones we test, Qwen3-0.6B and Qwen2.5-7B, and on Qwen2.5-7B it trails DAVID-GRPO by 9.1 average EM (Table 2).

Table 11: Qwen3-0.6B evaluation. EM and F1 of low-budget methods with the Qwen3-0.6B backbone. Best results are **bolded** and second-best underlined. *Search* is the average number of search actions per question over the six benchmarks. Qwen3-0.6B post-dates the 2024 collection of Antileak-m, so the Antileak-m scores are not contamination-free (Appendix D).

Method	HotpotQA		2wiki		Musique		Bamboogle		Bamtwoogle		Antileak-m		Avg.		
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	Search
Tree-GRPO	10.1	14.7	22.8	25.0	0.7	4.8	7.2	10.4	4.0	6.9	11.7	15.5	9.4	12.9	0.00
StepSearch	19.8	**30.9**	11.8	21.1	**8.7**	**16.8**	**23.2**	**34.7**	**17.0**	**20.7**	**35.4**	**45.0**	**19.3**	**28.2**	1.96
Search-R1-v0.3 *w/ retrieval reward*	9.7	14.6	**24.2**	27.3	1.0	5.4	5.6	8.8	4.0	7.9	12.5	15.8	9.5	13.3	0.01
DAVID-GRPO	**20.2**	28.9	22.7	**29.2**	5.9	11.9	17.6	26.0	13.0	17.6	34.5	41.9	19.0	25.9	**2.60**

Table 12: Ablation on evidence-guided continuation with Qwen2.5-7B. EM and F1 of the gold variant of DAVID-GRPO and its ablations under the low-budget setting. The variants follow Table 4.

Method	HotpotQA		2wiki		Musique		Bamboogle		Bamtwoogle		Antileak-m		Avg.	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
DAVID-GRPO	**35.2**	**45.6**	**37.7**	**44.5**	**14.2**	**21.6**	**40.0**	**52.3**	**44.0**	**47.1**	**55.2**	**59.8**	**37.7**	**45.1**
w/o continuation	31.9	41.6	31.2	37.9	12.4	19.8	32.0	40.6	41.0	44.2	52.5	57.0	33.5	40.2
w/o coverage reward	33.4	43.6	35.1	42.1	13.4	21.1	37.6	47.3	42.0	44.5	54.5	58.6	36.0	42.9

B.5 ABLATION ON QWEN2.5-7B

Table 12 repeats the ablation of Table 4 with Qwen2.5-7B for the gold variant. The full DAVID-GRPO system gives the best EM and F1 on all six benchmarks, and the two ablations fall below it by 4.2 and 1.7 average EM.

B.6 GENERALIZATION TO LONGER TOOL-USE CHAINS

Table 13: BFCL tool-use accuracy.

Method	Size	Acc. (%)
GPT-4.1-2025-04-14	-	43.0
Grok-4-0709	-	44.0
xLAM-2-1b-fc-r (reported)	1B	25.5
xLAM-2-1b-fc-r (64K override)	1B	41.5
+ GRPO	1B	42.5
+ DAVID-GRPO	**1B**	**44.5**

We also evaluate on the BFCL multi-turn long-context subset (Patil et al., 2025), where each problem requires 6.02 tool calls on average.

We use xLAM-2-1b-fc-r (Prabhakar et al., 2025) as the base model. Because it already supports multi-turn tool calling, we omit expert trajectory seeding and train with the same GRPO setup under a 64K context override. As shown in Table 13, DAVID-GRPO improves accuracy from 41.5 to 44.5 and outperforms standard GRPO by 2.0 pp. This suggests that DAVID-GRPO can help outside Wikipedia-style multi-hop QA.

B.7 TREE-GRPO UNDER A CONTROLLED SETTING

We check whether the shallow retrieval of Tree-GRPO in the low-budget setting is an artifact of our adaptation. First, the choice of the instruct checkpoint favors Tree-GRPO. With the Qwen2.5-1.5B base checkpoint, training fails, reaching 0.0 HotpotQA validation EM, whereas the instruct checkpoint reaches 14.5 (Table 14). Second, we retrain Tree-GRPO with the prompt, tag format, and training retrieval pool of DAVID-GRPO. As shown in Table 16, this controlled variant improves average EM from 10.7 to 11.6 and F1 from 14.8 to 17.0, but still issues almost no search actions (0.001 per question) and remains well below DAVID-GRPO. Third, shallow retrieval also appears in the publicly released high-budget runs of Tree-GRPO. In the 1.5B run, the number of valid searches per question stabilizes near 1.0 from step 49 onward, and all 23 readable responses in the final validation

Table 14: Tree-GRPO backbone choice. HotpotQA validation EM (%) of Tree-GRPO trained under the low-budget setting from the base and the instruct checkpoint of Qwen2.5-1.5B.

Tree-GRPO backbone	HotpotQA validation EM
Qwen2.5-1.5B	0.0
Qwen2.5-1.5B-Instruct	14.5

Table 15: Search counts of 7B agents on HotpotQA. Average number of search actions per question. The Tree-GRPO rows are the released high-budget run at three checkpoints; the DAVID-GRPO row is measured right after expert trajectory seeding, before the main GRPO stage.

7B agent	Checkpoint or stage	Searches
Tree-GRPO, high budget	Step 40 (minimum)	1.29
Tree-GRPO, high budget	Step 80 (maximum)	1.96
Tree-GRPO, high budget	Step 150 (best checkpoint)	1.82
DAVID-GRPO, low budget	After expert trajectory seeding	2.99

log perform exactly one search before answering. The 7B run averages 1.82 searches on HotpotQA at its best checkpoint and at most 1.96 at any checkpoint, whereas the low-budget DAVID-GRPO already averages 2.99 right after expert trajectory seeding, before the main GRPO stage (Table 15). Shallow retrieval of Tree-GRPO appears in every setting we examined, including the authors' own high-budget runs.

B.8 DYNAMICS OF PREFIX CONTINUATION

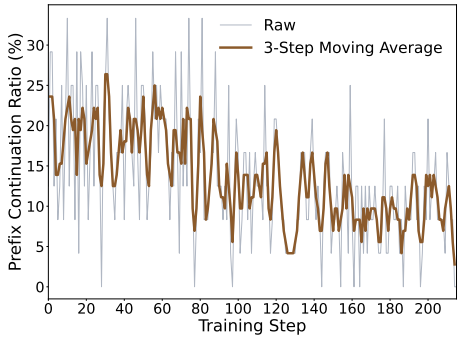


Figure 6: Prefix continuation ratio (%) over training steps based on Qwen2.5-1.5B with 4 NVIDIA RTX 3090 GPUs. The light gray line indicates the raw values, while the thick brown line denotes the 3-step moving average.

We investigate the temporal behavior of prefix continuation during training, as illustrated in Figure 6. The continuation ratio is higher in the early stages, between 20 and 33% of a batch (about 25% on the moving average), when the policy early in training rarely produces high-reward trajectories. As training progresses and the agent generates more high-reward trajectories on its own, the continuation frequency declines.

B.9 TRAINING STABILITY

We repeat the Qwen2.5-1.5B training three times to check seed sensitivity. We compare DAVID-GRPO against Search-R1-v0.3, the best low-budget baseline by average EM in Table 1. As shown in Table 17, the three-seed mean of average EM is 20.0 ± 1.7 for DAVID-GRPO against 9.8 ± 2.1 for Search-R1-v0.3, and DAVID-GRPO has higher mean EM on all six benchmarks.

Table 16: Tree-GRPO under a controlled setting. EM and F1 of Tree-GRPO with its released configuration and with the prompt, tag format, and training retrieval pool of DAVID-GRPO. The backbone is Qwen2.5-1.5B-Instruct for Tree-GRPO and Qwen2.5-1.5B for DAVID-GRPO.

Method	HotpotQA		2wiki		Musique		Bamboogle		Bamtwoogle		Antileak-m		Avg.	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Tree-GRPO (released configuration)	12.9	18.6	20.5	23.3	2.1	7.2	12.0	15.5	5.0	8.2	11.7	15.7	10.7	14.8
Tree-GRPO (controlled)	14.7	21.3	21.8	25.9	1.9	8.9	4.8	10.2	10.0	14.2	16.5	21.9	11.6	17.0
DAVID-GRPO	**24.8**	**33.8**	**27.2**	**32.3**	**7.1**	**12.6**	**14.4**	**24.2**	**22.0**	**25.4**	**36.3**	**41.1**	**22.0**	**28.2**

Table 17: Training stability. Mean and std. deviation of EM over three Qwen2.5-1.5B training runs.

Method	HotpotQA	2wiki	Musique	Bamboogle	Bamtwoogle	Antileak-m	Avg.
Search-R1-v0.3	14.5±5.0	18.0±2.8	2.4±1.4	5.9±1.8	4.0±1.0	13.8±3.5	9.8±2.1
DAVID-GRPO	**22.5**±2.1	**24.4**±2.5	**6.3**±1.0	**13.6**±2.9	**20.3**±2.1	**33.1**±3.0	**20.0**±1.7

B.10 ROBUSTNESS OF EVALUATION METRICS

EM and F1 are the standard metrics on these benchmarks, whose gold answers contain 2.45 words on average (1.69 to 2.80 per benchmark), and are used by Tree-GRPO (Ji et al., 2026), StepSearch (Zheng et al., 2025), and Search-R1-v0.3 (Jin et al., 2025a).

To verify that the gains of DAVID-GRPO are not specific to lexical overlap, we additionally evaluate all methods with LLM-as-judge accuracy, following prior multi-hop QA work (Sun et al., 2025; Song et al., 2025). We use Claude Sonnet 5 as the judge. Given the question, the gold answer, and the extracted predicted answer, the judge decides whether the prediction is semantically correct, with the prompt shown in Appendix F. As shown in Table 18, DAVID-GRPO remains the best-performing method on all six benchmarks, and the macro-average ranking of the low-budget methods is identical under EM, F1, and LLM-as-judge evaluation.

We also confirm that answer-extraction differences do not explain the gains. Table 19 reports final-answer extraction success rates aggregated across the six benchmarks. Extraction succeeds for at least 98% of the answers of every method, and the rate of DAVID-GRPO is slightly below the baselines, so the gains do not come from better compatibility with the extraction procedure.

Table 19: Final-answer extraction success rate (%) aggregated across the six benchmarks under the low training budget setting.

Method	Extraction success (%)
Tree-GRPO	**99.94**
StepSearch	99.90
Search-R1-v0.3	99.82
DAVID-GRPO	98.27

B.11 NO-TRAINING BASELINES

We report no-training controls to verify that the gains of DAVID-GRPO do not come from the tool-use prompt scaffold, the additional retrieval context, or the existing capability of the base model. We evaluate two settings with Qwen2.5-1.5B on all six benchmarks under the same overall experimental setup as DAVID-GRPO.

- **Base Agent w/o RL**: the untrained model uses the same agent prompt and search interface as DAVID-GRPO.
- **Fixed Retrieval w/o Training**: the question is used as the retrieval query, and the retrieved passages are provided as context without agentic search.

As shown in Table 20, both settings achieve at most 0.1 EM on every benchmark. This result is consistent with the no-seeding baseline in Figure 3, which uses the same agent prompt and obtains zero EM on all benchmarks.

Table 18: LLM-as-judge evaluation. LLM-as-judge accuracy (%) across six multi-hop QA benchmarks under the low training budget setting. The backbone model is Qwen2.5-1.5B except for Tree-GRPO with Qwen2.5-1.5B-Inst.

Method	HotpotQA	2wiki	Musique	Bamboogle	Bamtwoogle	Antileak-m	Avg.
Tree-GRPO	21.1	23.8	5.8	15.2	7.0	12.3	14.2
StepSearch	21.9	18.4	5.8	8.0	7.0	14.5	12.6
Search-R1-v0.3	31.4	27.2	9.1	15.2	8.0	19.8	18.5
DAVID-GRPO	**40.6**	**34.7**	**13.4**	**28.8**	**28.0**	**40.9**	**31.1**

Table 20: No-training baselines. EM (%) of Qwen2.5-1.5B under two no-training settings across six multi-hop QA benchmarks, using the same overall experimental setup as DAVID-GRPO.

Setting	HotpotQA	2wiki	Musique	Bamboogle	Bamtwoogle	Antileak-m
Base Agent w/o RL	0.0	0.1	0.0	0.0	0.0	0.0
Fixed Retrieval w/o Training	0.0	0.1	0.0	0.0	0.0	0.0

B.12 RETRIEVAL BEHAVIOR ANALYSIS

Table 21 splits the document recall of Table 3 into bridge and answer documents on the benchmarks that annotate them. DAVID-GRPO has the highest recall in every column, and Tree-GRPO has zero because it issues no search actions.

§5.1 discusses the share of searches issued after the support set is complete (*Unnec.* in Table 3).

Table 22 separates the two components of evidence-guided continuation at the level of retrieval behavior. Adding the evidence-coverage reward to the outcome-only variant raises complete-evidence recall (*All*) from 18.5 to 27.4 on HotpotQA and from 4.8 to 5.5 on MuSiQue, and adding prefix continuation raises it further to 31.6 and 8.2. EM follows the same order. Neither component consistently changes document precision, and the increase in searches issued after the evidence chain is complete stays within 2.9 points for the coverage reward and 0.5 points for prefix continuation.

C BENCHMARK STATISTICS

D KNOWLEDGE CUTOFF DATES

To verify the validity of our evaluation on **AntiLeakBench** (Wu et al., 2025), we report the official knowledge cutoff dates for the small language agents used in our experiments. As shown in Table 24, the Qwen2.5 and Llama-3.2 models have a knowledge cutoff of December 2023. Since the multi-hop tasks in AntiLeakBench are constructed from data generated in 2024 (Table 23), the test samples were not available during pre-training. Qwen3-0.6B is the exception: its knowledge cutoff is not disclosed and it was released in April 2025, after the collection period, so we do not treat its Antileak-m scores (Table 11) as contamination-free.

E IMPLEMENTATION DETAILS

E.1 TRAINING AND EVALUATION SETUP

As described in §4.1, both training and evaluation were performed on a system equipped with four NVIDIA RTX 3090 GPUs (24GB VRAM each), except that the Qwen2.5-7B run was trained on eight NVIDIA RTX A6000 GPUs (48GB VRAM each) with the same batch size and rollout budget.

Training. The default model trains on the 4,000 constructed instances of §4.1. The gold variant and the analyses in Appendix B use a subset of 5,276 examples from the HotpotQA (Yang et al., 2018) training set, provided by Zhang et al. (2025). The retrieval infrastructure uses `E5-base-v2` (Wang et al., 2022) as the dense encoder and FAISS (Johnson et al., 2019) for efficient indexing and similarity search.

Table 21: Document recall (%; ↑) of cumulative retrieved documents $\hat{\mathcal{D}}_{union}$ (Eq. (2)). Any and All measure coverage of annotated answer/bridge documents in D. Search-R1-v0.3 is trained with its retrieval reward. Base model is Qwen2.5-1.5B.

Method	HotpotQA				Musique			
	Bridge-doc		Answer-doc		Bridge-doc		Answer-doc	
	Any	All	Any	All	Any	All	Any	All
Tree-GRPO	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
StepSearch	50.6	48.1	38.0	26.6	28.1	18.4	11.9	9.5
Search-R1-v0.3	26.3	25.1	19.8	14.0	17.2	10.4	6.4	5.0
DAVID-GRPO	**59.1**	**58.0**	**51.1**	**42.4**	**48.5**	**27.5**	**18.5**	**15.8**

Table 22: Retrieval behavior of the ablations. Retrieval statistics and EM (%) on HotpotQA and MuSiQue with the Qwen2.5-1.5B backbone. Each row adds one component to the row above. The metrics follow Table 3.

Dataset	Reward and continuation	Search	Prec.	Any	All	Unnec.	EM
HotpotQA	Outcome reward only (*w/o coverage reward*)	2.52	29.2	42.9	18.5	6.4	20.8
	+ Evidence-coverage reward (*w/o continuation*)	2.60	**30.8**	49.7	27.4	9.3	23.3
	+ Prefix continuation (DAVID-GRPO)	3.05	28.2	**52.3**	**31.6**	9.8	**24.8**
MuSiQue	Outcome reward only (*w/o coverage reward*)	2.86	**14.6**	22.5	4.8	1.9	4.5
	+ Evidence-coverage reward (*w/o continuation*)	3.05	13.7	24.5	5.5	2.1	5.8
	+ Prefix continuation (DAVID-GRPO)	3.48	14.3	**25.9**	**8.2**	2.6	**7.1**

Evaluation. Our evaluation pipeline uses FlashRAG (Jin et al., 2025c), following the architectures used in Search-R1(-v0.3) (Jin et al., 2025b;a) and Tree-GRPO (Ji et al., 2026). The retrieval and corpus configuration has two settings to align the corpus date with the datasets:

Pipeline A FlashRAG-based retrieval-augmented generation pipeline.

Retriever The Search-R1 retriever, employing the same `E5-base-v2` and FAISS setup as described in the training phase.

Corpus (default) For the majority of tasks, we use the `wiki-18.jsonl` corpus, following the standard setup in multi-hop reasoning evaluations.

Corpus (2024) To accommodate the 2024 subset of AntiLeakBench, we constructed a dedicated index using the Wikipedia 2024 dump. This lets the retriever access information contemporary to the evaluation queries, while keeping the retriever model identical to the default setting.

E.2 EXISTING ASSETS AND LICENSES

Table 26 summarizes the main external datasets, models, and software used in our experiments. When a public source does not specify a license, we mark it as not specified.

E.3 EXPERT TRAJECTORY SEEDING

To build the expert trajectory set, we use the HotpotQA trajectory annotations from AutoCoA (Zhang et al., 2025), where the expert policy π^* is DeepSeek-R1-Distill-Qwen-32B (Guo et al., 2025). We randomly selected $k{=}4$ trajectories returning correct answers as the set X_{exp}. The selected examples are shown in Appendix H. For efficiency, we skip prefix continuation in the seeding phase.

Sensitivity to the Expert Trajectory Set. To check whether the results depend on the specific four trajectories, we sample two additional sets of four expert trajectories independently and retrain DAVID-GRPO with the same setup. Table 27 shows average EM of 22.0, 24.3, and 22.3 for the three sets (22.9±1.3). Every set outperforms Search-R1-v0.3 on all six benchmarks. The largest variation is on Bamboogle and Antileak-m, which contain 125 and 455 questions.

Table 23: Statistics of the multi-hop QA evaluation benchmarks.

Name	Size	Comment
HotpotQA (Yang et al., 2018)	7,405	Used in Search-R1, Tree-GRPO, etc.
2WikiMultiHopQA (Ho et al., 2020)	12,576	
MuSiQue Trivedi et al. (2022)	2,417	
Bamboogle (Press et al., 2023)	125	
BamTwoogle (Aksitov et al., 2024)	100	Advanced version of Bamboogle
AntiLeakBench (multi-hop; Wu et al., 2025)	455	2024 subset

Table 24: Knowledge cutoff dates for the language agents evaluated in this work. The Qwen2.5 and Llama-3.2 models pre-date the data collection period of AntiLeakBench (2024); Qwen3-0.6B does not.

Model Family	Specific Model	Knowledge Cutoff
Qwen2.5	Qwen2.5-1.5B Qwen2.5-1.5B-Instruct Qwen2.5-0.5B Qwen2.5-0.5B-Instruct Qwen2.5-7B	Dec 2023[†]
Llama 3.2	Llama-3.2-1B-Instruct	Dec 2023[‡]
Qwen3	Qwen3-0.6B	Not disclosed; released Apr 2025[§]

[†] https://github.com/QwenLM/Qwen3/issues/525
[‡] https://huggingface.co/meta-llama/Llama-3.2-1B
[§] https://github.com/QwenLM/Qwen3

Table 27: Sensitivity to the expert trajectory set. EM (%) of DAVID-GRPO trained with three independently sampled sets of four expert trajectories on Qwen2.5-1.5B, with the best low-budget baseline for reference. Set 1 is the default set used in the main experiments.

Method	HotpotQA	2wiki	Musique	Bamboogle	Bamtwoogle	Antileak-m	Avg.
Search-R1-v0.3	19.0	21.3	3.6	8.0	3.0	16.7	11.9
DAVID-GRPO (set 1)	24.8	**27.2**	7.1	14.4	22.0	36.3	22.0
DAVID-GRPO (set 2)	**26.0**	26.9	7.7	**22.4**	**23.0**	**39.6**	**24.3**
DAVID-GRPO (set 3)	24.9	26.4	**8.5**	**22.4**	20.0	31.4	22.3
Mean ± std.	25.2±0.7	26.8±0.4	7.8±0.7	19.7±4.6	21.7±1.5	35.8±4.1	22.9±1.3

For the expert trajectory ratio, we keep the standard policy ratio $\pi_\theta(\tau^*)/\pi_{\theta_{old}}(\tau^*)$ and set the additional expert-policy correction $\omega^*(\theta)$=1. Table 28 compares different choices of $\omega^*(\theta)$ in expert trajectory seeding. The exact trajectory-level correction yields 0 EM because the policy early in training assigns near-zero probability to expert trajectories, and the action-level correction gives lower EM than the constant correction.

Table 28: Off-policy correction ablation. EM results for Qwen2.5-1.5B with different choices of the expert-policy correction $\omega^*(\theta)$ in expert trajectory seeding.

$\omega^*(\theta)$	HotpotQA	2wiki	Musique	Bamboogle	Bamtwoogle	Antileak-m	Avg.
$\pi_\theta(\tau^*)/\pi^*(\tau^*)$	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\pi_\theta(a_t^* \mid s_t^*)/\pi^*(a_t^* \mid s_t^*)$	14.6	18.1	3.1	10.4	9.0	24.4	13.3
1	**24.8**	**27.2**	**7.1**	**14.4**	**22.0**	**36.3**	**22.0**

Table 29 compares DAVID-GRPO under different annotation budgets. Increasing the number of expert trajectories from k=4 to k=96 improves average EM from 22.0 to 24.2, while using the full AutoCoA trajectory set with SFT initialization (k=5,276) further improves average EM to 30.8. We

Table 25: Hyperparameters for Training and Evaluation.

Parameter	Expert Trajectory Seeding	Main GRPO	Evaluation
Dataset Size	4 (expert trajectories)	4,000	-
Batch Size	4	24	128
Rollouts	5	5–10 (avg. 6.1)	1
Group Size (G)	5	5	-
Max Search	5	5	-
Learning Rate	1×10^{-5}	1×10^{-6}	-
KL Loss Coefficient (β)	1×10^{-3}	1×10^{-3}	-
Clipping Parameter (ϵ)	0.2	0.2	-
Training Steps	50	215	-
Max Prompt Len.	1,024	1,024	2,048
Max Response Len.	8,192	8,192	12,288
Temperature	0.6	0.6	0.6
Top-p	1.0	1.0	1.0
Top-k	-1	-1	-1

Table 26: Existing assets used in the experiments and their public licenses or terms.

Asset	Use in this paper	License / terms
HotpotQA (Yang et al., 2018)	Training and evaluation data	CC BY-SA 4.0
2WikiMultiHopQA (Ho et al., 2020)	Evaluation data	Apache-2.0
MuSiQue (Trivedi et al., 2022)	Evaluation data	Not specified by public source
Bamboogle (Press et al., 2023) and BamTwoogle (Aksitov et al., 2024)	Evaluation data	Not specified by public source
AntiLeakBench (Wu et al., 2025)	Evaluation data	Not specified by public source
BFCL (Patil et al., 2025)	Tool-use evaluation data	Apache-2.0
Qwen2.5 / Qwen3 models (Qwen et al., 2025; Yang et al., 2025)	Base language agents	Apache-2.0
Llama-3.2-1B (Grattafiori et al., 2024)	Base language agent	Llama 3.2 Community License
xLAM-2-1b-fc-r (Prabhakar et al., 2025)	Base tool-use agent	Llama 3 Community License
`E5-base-v2` (Wang et al., 2022)	Dense retriever encoder	MIT
FAISS (Johnson et al., 2019)	Vector index and retrieval	MIT
FlashRAG (Jin et al., 2025c)	Evaluation infrastructure	MIT

keep k=4 as the default setting because the main goal is to test whether expert trajectory seeding can stabilize low-budget RL with only a handful of expert trajectories.

Table 29: Effect of annotation budget. EM results on Qwen2.5-1.5B with different numbers of expert trajectory annotations.

Setting	K	HotpotQA	2wiki	Musique	Bamboogle	Bamtwoogle	Antileak-m	Avg.
DAVID-GRPO	4	24.8	**27.2**	7.1	14.4	22.0	36.3	22.0
DAVID-GRPO	96	27.6	26.9	8.6	20.8	22.0	39.1	24.2
DAVID-GRPO with Full SFT	5,276	**32.6**	**28.1**	**12.2**	**33.6**	**32.0**	**46.2**	**30.8**

Controlled RL Variants. To isolate the contribution beyond the expert trajectory seeding component of DAVID-GRPO, we retain this component and replace the subsequent RL method with vanilla GRPO (Shao et al., 2024), StepSearch (Zheng et al., 2025), Search-R1-v0.3 (Jin et al., 2025a), or Tree-GRPO (Ji et al., 2026). All variants use the same four expert trajectories and low-budget setup. Table 30 shows that DAVID-GRPO achieves the highest average EM of 22.0, exceeding the Tree-GRPO variant by 1.6 points. Access to the expert trajectories alone does not explain the gain.

Table 30: EM (%) for controlled variants of Qwen2.5-1.5B that retain the expert trajectory seeding component of DAVID-GRPO. All variants use the same four expert trajectories and low-budget setup. Best results are **bolded**; second-best results are underlined.

Method	HotpotQA	2wiki	Musique	Bamboogle	Bamtwoogle	Antileak-m	Avg.
Expert trajectory seeding + Vanilla GRPO	20.8	22.0	4.5	10.4	10.0	31.4	16.5
Expert trajectory seeding + StepSearch	20.4	19.8	3.6	9.6	9.0	19.1	13.6
Expert trajectory seeding + Search-R1-v0.3	22.7	23.3	6.6	13.6	19.0	28.8	19.0
Expert trajectory seeding + Tree-GRPO	23.6	23.6	**7.5**	12.8	**22.0**	33.2	20.4
DAVID-GRPO	**24.8**	**27.2**	7.1	**14.4**	**22.0**	**36.3**	**22.0**

E.4 EVIDENCE-COVERAGE REWARD

Table 31 compares the proposed *evidence-coverage reward* (r_g) with retrieval rewards from prior work. Starting from the same expert trajectory seeding checkpoint, we train with three reward signals while keeping the other settings fixed. DAVID-GRPO (Coverage) gives the best results on most benchmarks. Search-R1-v0.3, which rewards answer-document retrieval, is second, while StepSearch, which uses step-wise lexical similarity, is lower in this setup. On Antileak-m, DAVID-GRPO improves over Search-R1-v0.3 by 4.4 pp in EM and 5.4 pp in F1.

Table 31: Comparison of Retrieval Rewards. Multi-hop QA results trained on Qwen2.5-1.5B.

Retrieval Reward	HotpotQA		2wiki		Musique		Bamboogle		Bamtwoogle		Antileak-m		Avg.	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
StepSearch	20.4	29.0	19.8	24.8	3.6	8.2	9.6	18.2	9.0	12.0	19.1	23.1	13.6	19.2
Search-R1-v0.3	22.7	31.0	23.3	28.0	**6.6**	**11.4**	13.6	**22.8**	**19.0**	23.4	28.8	32.8	19.0	24.9
DAVID-GRPO (Coverage)	**23.3**	**31.9**	**24.4**	**30.0**	5.8	11.3	**14.4**	20.9	**19.0**	**23.8**	**33.2**	**38.2**	**20.0**	**26.0**

To examine whether explicitly accounting for retrieval precision improves the evidence-coverage reward, we replace our recall-oriented r_g with an F1-oriented variant that equally weights the precision and recall of $\hat{\mathcal{D}}_{union}$ against D. As shown in Table 32, the F1-oriented reward reduces EM on all six benchmarks. Table 33 explains this drop. Incorporating precision weakens the reward for additional searches because they may retrieve documents outside D, even when required evidence remains missing. This suppresses retrieval, reducing the average number of searches from 3.05 to 1.23 on HotpotQA and from 3.48 to 1.49 on MuSiQue, and consequently lowers both document recall metrics and QA performance (see also Table 3).

Table 32: Recall- and F1-oriented retrieval rewards. EM (%) when training with the recall-oriented evidence-coverage reward of DAVID-GRPO and with an F1-oriented variant that equally weights precision and recall. The backbone model is Qwen2.5-1.5B.

Retrieval reward	HotpotQA	2wiki	Musique	Bamboogle	Bamtwoogle	Antileak-m	Avg.
Recall-oriented (DAVID-GRPO)	**24.8**	**27.2**	**7.1**	**14.4**	**22.0**	**36.3**	**22.0**
F1-oriented	19.6	17.7	4.8	8.8	14.0	24.2	14.8

Table 33: Retrieval behavior of recall- and F1-oriented rewards. Retrieval statistics and EM (%) on HotpotQA and MuSiQue with the Qwen2.5-1.5B backbone. The metrics follow Table 3.

Dataset	Retrieval reward	Search	Prec.	Any	All	Unnec.	EM
HotpotQA	Recall-oriented (DAVID-GRPO)	3.05	28.2	**52.3**	**31.6**	9.8	**24.8**
	F1-oriented	1.23	**39.1**	41.5	15.1	1.4	19.6
MuSiQue	Recall-oriented (DAVID-GRPO)	3.48	14.3	**25.9**	**8.2**	2.6	**7.1**
	F1-oriented	1.49	**18.0**	18.4	2.8	0.4	4.8

A related concern is the case where the retriever fails to return any document of the support set, since r_g depends on exact membership in D. We measured this over a full training run of DAVID-GRPO.

Only 0.88% of the sampled groups contained no trajectory that retrieved a support document. In the remaining 99.12%, at least one trajectory retrieved a support document, so the group-relative advantage carried an evidence signal. In the rare all-miss groups, r_g is zero for every trajectory and the update falls back to the outcome reward.

We set λ=0.5 in the main experiments, giving equal weight to the evidence-coverage reward r_g and the outcome reward r_o. Table 34 varies λ; although the best value differs by dataset, λ=0.5 achieves the highest average EM among the tested values.

Table 34: EM results for Qwen2.5-1.5B with different weights on the evidence-coverage reward.

λ	HotpotQA	2wiki	Musique	Bamboogle	Bamtwoogle	Antileak-m	Avg.
0.25	21.4	23.2	6.1	16.8	18.0	32.3	19.6
0.50	**24.8**	**27.2**	**7.1**	14.4	**22.0**	**36.3**	**22.0**
0.75	21.3	24.1	5.5	**17.6**	18.0	35.2	20.3

E.5 Prefix Continuation

To efficiently generate the refined trajectory τ'_M, we employ a rejection resampling strategy. Given the truncated history $\tau_{M,\leq t'}$ (where t' is the last step with valid retrieval), we sample l independent candidate completions $\{\tilde{\tau}^{(1)}, \ldots, \tilde{\tau}^{(l)}\}$ from the policy π_θ using temperature sampling. We then evaluate the total reward for each completion and select the one with the highest score as the final continued trajectory τ'_M:

$$\tau'_M = \underset{\tilde{\tau} \in \{\tilde{\tau}^{(1)}, \ldots, \tilde{\tau}^{(l)}\}}{\arg\max} R(\tilde{\tau}). \tag{7}$$

In our experiments, we set the number of continuation samples l=5.

E.6 Details of Corpus-Graph Evidence Construction

§4.1 summarizes the construction pipeline. Following the graph-based construction procedure of HotpotQA (Yang et al., 2018), we use the Wikipedia dump released with HotpotQA as follows:

1. Build a hyperlink graph from the Wikipedia documents.
2. Sample subgraphs according to the target hop distribution. Among the 4,000 examples, 55%, 44%, 0.5%, and 0.5% are 2-, 3-, 4-, and 5-hop, respectively.
3. Extract entities and inter-document relations from each subgraph using the two-shot prompt in Appendix F.1. We generate a question–answer pair from the leaf document and compose the multi-hop question backward using the one-shot prompt. We use Qwen2.5-3B (Qwen et al., 2025), the smallest model that produced stable generations. Qwen2.5-1.5B produced degenerate outputs.
4. Retain an example only if (i) it follows the required format, (ii) its support documents do not occur in the top-20 BM25 results for the initial question, (iii) the answer cannot be produced without the support documents, (iv) no proper subset of the support documents is sufficient, and (v) the generated answer is correct.

The support set D is the document chain of the sampled subgraph. The pipeline requires the source corpus but does not use manually annotated support sets.

F PROMPTS

DAVID-GRPO

[Prompt]

```
You are an intelligent reasoning assistant with access to a web
    search tool to help you obtain accurate information. When
    answering the user's questions, please adhere to the following
    guidelines:

1. Uncertainty Handling:
   - Encourage the use of web search to solve the problems mentioned
     during reasoning.
   - If you are uncertain or do not have sufficient knowledge about
    a specific detail, you must perform a web search.
   - To perform a web search the format must be in the format
    mentioned in point 2 Web Search Format
   - You can perform multiple web searches to ensure that the answer
     is correct.
   - The web search should be initiated only within your internal
    reasoning (the "thinking" phase), and it should not appear in
    your final answer to the user.

2. Web Search Format:
   - When calling the web search tool, use the following exact
    format:
     <begin_search> your query here </end_search>
   - After you perform a web search, you will receive results
    enclosed in:
     <search_result> ... </search_result>
   - You can then use this information to further refine your
    reasoning and answer.

3. Process:
   - If you encounter any uncertainty or unknown knowledge, embed a
    web search query within your internal thoughts (surrounded by
    <think></think> tags) using the specified format.
   - Incorporate the search results into your reasoning before
    finalizing your answer.
   - You should provide your final answer in the format \boxed{
    YOUR_ANSWER}.

Now, please answer the user's question below:
{question}
```

LLM-as-Judge

[Prompt]

```
Given a Question and its Golden Answer, verify whether the Predicted
     Answer is correct. The prediction is correct if it fully aligns
     with the meaning and key information of the Golden Answer.
    Respond with True if the prediction is correct and False
    otherwise.

Question: {question}
Golden Answer: {gold_answer}
```

```
Predicted Answer: {predicted_answer}
```

F.1 Synthetic Data Generation

Prompt A extracts a referring expression for an inter-document relation, and Prompt B generates a leaf-document question–answer pair. The referring expressions are substituted backward along the sampled chain to form a multi-hop question.

Prompt A: Entity and Relation Extraction

[Prompt]

```
System: "You rewrite Wikipedia facts into short referring
    expressions. You never add information that is not in the given
    paragraph."

Write a short noun phrase that uniquely identifies the target entity
     using ONLY facts stated in the paragraph.

Rules:
- Write the literal token <SOURCE> wherever the paragraph's title
    belongs. Never spell the title out.
- The phrase MUST NOT contain the target entity's name or any word
    from it.
- Use a definite noun phrase: "the director of <SOURCE>", "the town
    where <SOURCE> was founded".
- One line, no quotes, no explanation.

### Example 1
Paragraph (title: One Night Stand (1984 film)):
One Night Stand is a 1984 film directed by John Duigan.
Target entity: John Duigan
Phrase: the director of <SOURCE>

### Example 2
Paragraph (title: Then We Came to the End):
Then We Came to the End is a 2007 novel by Joshua Ferris, published
    by Little, Brown and Company.
Target entity: Little, Brown and Company
Phrase: the publisher of <SOURCE>

### Response
Paragraph (title: {src_title}): {src_text}
Target entity: {dst_title}
Phrase:
```

Prompt B: Question and Answer Generation

[Prompt]

```
System: "You write single-hop factual questions grounded in one
    Wikipedia paragraph. You copy answers verbatim from the
    paragraph."

Write ONE short factual question about the given title that the
    paragraph answers.
```

```
Rules:
- Write the literal token <TITLE> wherever the title belongs. Never
    spell it out.
- The answer MUST be COPIED CHARACTER FOR CHARACTER from the
    paragraph, at most 5 words.
- Prefer a name, a date, or a number as the answer -- not a
    descriptive phrase.
- The answer MUST NOT appear anywhere in the question.
- The question's wh-word must match the answer: "How many"/"How long
    " need a number, "When" needs a date, "Who" needs a person.
- Ask about an attribute, never about the title's own name.

### Example
Paragraph (title: Little, Brown and Company):
Little, Brown and Company is an American publisher founded in 1837
    by Charles Little and James Brown.
{"question": "In what year was <TITLE> founded?", "answer": "1837",
    "answer_type": "date"}

### Response
Paragraph (title: {title}): {text}
Return JSON only:
{"question": "...", "answer": "...", "answer_type": "entity|date|
    number|span"}
```

G CASE STUDY

G.1 DAVID-GRPO

DAVID-GRPO (Correct)

Level: Hard **Source:** HotpotQA **Reasoning Type:** Comparison

[Question]

```
Are the Laleli Mosque and Esma Sultan Mansion located in the same
    neighborhood?
```

[Gold Answer]

```
No
```

[Titles of Support Evidence Documents (D)]

```
Laleli Mosque, Esma Sultan Mansion
```

[Response]

```
<think>
Okay, I need to determine if the Laleli Mosque and Esma Sultan
    Mansion are located in the same neighborhood. Let me start by
    recalling what I know about these two landmarks.

First, I remember that I learned in history class that the Laleli
    Mosque is an important Islamic structure in Istanbul. I'll try
    to remember if it's part of a larger complex or if it's located
    in a specific neighborhood. I'll search for "Laleli Mosque
    location Istanbul" to get the basic information.

<begin_search>Laleli Mosque location Istanbul</end_search>

<search_result> result 1: "Laleli Mosque"
Laleli Mosque The Laleli Mosque () is an 18th-century Ottoman
    imperial mosque located in Laleli, Fatih, Istanbul, Turkey. The
    Laleli Mosque was built by Sultan Mustafa III from 1760-1763,
    designed in the baroque style by Ottoman imperial architect
    Mehmet Tahir Ağa. The complex was destroyed by a fire in 1783
    shortly after its completion and was immediately rebuilt. A fire
     in 1911 destroyed the madrasah, and subsequent road
    construction work destroyed many other auxiliary structures to
    the mosque. The mosque was built on a high terrace over a
    complex of vaulted shops, whose rents were intended to
    financially support the
result 2: "Tekeli Mehmet Pasha Mosque"
Tekeli Mehmet Pasha Mosque The Tekeli Mehmet Paşa Mosque () is a
    mosque in the city of Antalya, Turkey. Mosque takes its name
    from Lala Mehmed Pasa. Built in the 18th century in the
    Kalekapisi district, the mosque is one of the most important
    Ottoman mosques in the city. The main dome, which rises on a
```

high rim, is supported by three semi domes, one each in the east, west, and south directions, as well as three domes on the northern side. There are tiled panels decorated with Koranic verses in taliq script on the pointed arch-formed lunettes of the
result 3: "Tekeli Mehmet Pasha Mosque"
windows on the northern façade of the mosque and inside it. Tekeli Mehmet Pasha Mosque The Tekeli Mehmet Paşa Mosque () is a mosque in the city of Antalya, Turkey. Mosque takes its name from Lala Mehmed Pasa. Built in the 18th century in the Kalekapisi district, the mosque is one of the most important Ottoman mosques in the city. The main dome, which rises on a high rim, is supported by three semi domes, one each in the east, west, and south directions, as well as three domes on the northern side. There are tiled panels decorated with Koranic
</search_result>

Alright, the first result clearly mentions that the Laleli Mosque is located in Laleli, Fatih, Istanbul, Turkey. Now, I need to find out about the Esma Sultan Mansion to see if they are in the same neighborhood.

I'll search for "Esma Sultan Mansion location Istanbul" to get that information.

<begin_search>Esma Sultan Mansion location Istanbul</end_search>

<search_result> result 1: "Esma Sultan Mansion"
Esma Sultan Mansion The Esma Sultan Mansion (), a historical yalı () located at Bosphorus in Ortaköy neighborhood of Istanbul, Turkey and named after its original owner Esma Sultan, is used today as a cultural center after being redeveloped. The three-storey brick building was designed by the renowned architect Sarkis Balyan and finished in 1875 next to Ortaköy Mosque. It was presented to Esma Sultan, the daughter of Ottoman Sultan Abdulaziz, as a wedding gift. The mansion remained in the possession of the Ottoman dynasty until 1915. The building was subsequently used first as a tobacco warehouse and then as
result 2: "Esma Sultan Mansion"
guests in the ground floor and 600 guests in the first floor. The venue also hosts concerts of the Istanbul International Jazz Festival and the Istanbul International Music Festival. Esma Sultan Mansion The Esma Sultan Mansion (), a historical yalı () located at Bosphorus in Ortaköy neighborhood of Istanbul, Turkey and named after its original owner Esma Sultan, is used today as a cultural center after being redeveloped. The three-storey brick building was designed by the renowned architect Sarkis Balyan and finished in 1875 next to Ortaköy Mosque. It was presented to Esma Sultan, the daughter of Ottoman Sultan Abdulaziz,
result 3: "Esma Sultan Mansion"
a coal depot from 1920 until 1975, when it was destroyed by a fire. The ruin, consisting only of the outer walls of the building, was purchased in the early 1990s by The Marmara hotel chain. Following a renovation with additions designed by architects Haluk Sezgin and Philippe Robert, the mansion was opened in 2001 as a multipurpose event venue. Within the brick exterior, which was left as the original, a steel and glass structure is incorporated. According to the GAD Architecture group, a

redesign by the architect Gökhan Avcıoğlu was completed in 2005. The building includes a bar, a </search_result>

From the first result, the Esma Sultan Mansion is located at Ortaköy neighborhood of Istanbul. Now, I need to check if Laleli is part of the same neighborhood. I'll search for "Laleli neighborhood Istanbul" to confirm its location.

<begin_search>Laleli neighborhood Istanbul</end_search>

<search_result> result 1: "Laleli Mosque"
Laleli Mosque The Laleli Mosque () is an 18th-century Ottoman imperial mosque located in Laleli, Fatih, Istanbul, Turkey. The Laleli Mosque was built by Sultan Mustafa III from 1760-1763, designed in the baroque style by Ottoman imperial architect Mehmet Tahir Ağa. The complex was destroyed by a fire in 1783 shortly after its completion and was immediately rebuilt. A fire in 1911 destroyed the madrasah, and subsequent road construction work destroyed many other auxiliary structures to the mosque. The mosque was built on a high terrace over a complex of vaulted shops, whose rents were intended to financially support the
result 2: Etiler
Etiler Etiler is a neighbourhood on the European side of Istanbul, Turkey, in the district of Beşiktaş, close to the business quarters of Levent and Maslak. Etiler is famous for its upmarket cafés, pubs, night clubs, restaurants, gyms, coiffeurs, fashion shops and shopping malls, such as Akmerkez. It is a favourite area among Istanbul's elite. The quarter also has many villas and private residences. The name ""Etiler"" is an older name in Turkish for the ""Hittites"", as it was fashionable in the early years of the Turkish Republic to give the names of ancient Anatolian civilizations to the new districts
result 3: "Şişli"
Şişli Şişli () is one of 39 districts of Istanbul, Turkey. Located on the European side of the city, it is bordered by Beşiktaş to the east, Sarıyer to the north, Eyüp and Kağıthane to the west, and Beyoğlu to the south. In 2009, Şişli had a population of 316,058. Until the 1800s, Şişli was open countryside, used for hunting and agriculture and leisure. It was developed as a middle class residential district during the last years of the Ottoman Empire and the early years of the Turkish Republic (the late 19th-early 20th centuries). French culture was an important influence </search_result>

The first result clearly states that the Laleli Mosque is located in Laleli, Fatih, Istanbul, Turkey. It's located in the Fatih district, specifically in the Laleli neighborhood. Therefore, Laleli is part of the Fatih neighborhood, not Ortaköy.

Since the Esma Sultan Mansion is in Ortaköy, and the Laleli Mosque is in Laleli, they are indeed located in different neighborhoods.

</think>

The Laleli Mosque is in the Fatih neighborhood, and the Esma Sultan Mansion is in the Ortaköy neighborhood. Thus, they are not located in the same neighborhood.

```
\boxed{No}
```

G.2 TREE-GRPO

Tree-GRPO (Incorrect)

Level: Hard **Source:** HotpotQA **Reasoning Type:** Comparison

[Question]

```
Are the Laleli Mosque and Esma Sultan Mansion located in the same
    neighborhood?
```

[Gold Answer]

```
No
```

[Titles of Support Evidence Documents (D)]

```
Laleli Mosque, Esma Sultan Mansion
```

[Response]

```
<think>
The Laleli Mosque is located in Istanbul, Turkey, which is the
    capital city of the country.
Esma Sultan Mansion is located in East Fourth district, Beyoglu
    neighborhood, Istanbul, Turkey.
Both are located in Istanbul, Turkey.
</think>
<answer>Yes</answer>
```

G.3 STEPSEARCH

StepSearch (Incorrect)

Level: Hard **Source:** HotpotQA **Reasoning Type:** Comparison

[Question]

```
Are the Laleli Mosque and Esma Sultan Mansion located in the same
    neighborhood?
```

[Gold Answer]

No

[Titles of Support Evidence Documents (D)]

Laleli Mosque, Esma Sultan Mansion

[Response]

<answer> no. <
<warning> My previous action is invalid. If I want to search, I should put the query between **<search>** and **</search>**. If I want to give the final answer, I should put the answer between **<answer>** and **</answer>**. Let me try again. </warning>
<search> Are the Laleli Mosque and Esma Sultan Mansion located in the same neighborhood **</search>**

<information> Doc 1<## Title: ”Esma Sultan Mansion” ##> Esma Sultan Mansion The Esma Sultan Mansion (), a historical yalı () located at Bosphorus in Ortaköy neighborhood of Istanbul, Turkey and named after its original owner Esma Sultan, is used today as a cultural center after being redeveloped. The three-storey brick building was designed by the renowned architect Sarkis Balyan and finished in 1875 next to Ortaköy Mosque. It was presented to Esma Sultan, the daughter of Ottoman Sultan Abdulaziz, as a wedding gift. The mansion remained in the possession of the Ottoman dynasty until 1915. The building was subsequently used first as a tobacco warehouse and then as
Doc 2<## Title: ”Esma Sultan Mansion” ##> guests in the ground floor and 600 guests in the first floor. The venue also hosts concerts of the Istanbul International Jazz Festival and the Istanbul International Music Festival. Esma Sultan Mansion The Esma Sultan Mansion (), a historical yalı () located at Bosphorus in Ortaköy neighborhood of Istanbul, Turkey and named after its original owner Esma Sultan, is used today as a cultural center after being redeveloped. The three-storey brick building was designed by the renowned architect Sarkis Balyan and finished in 1875 next to Ortaköy Mosque. It was presented to Esma Sultan, the daughter of Ottoman Sultan Abdulaziz,
Doc 3<## Title: ”Laleli Mosque” ##> Laleli Mosque The Laleli Mosque () is an 18th-century Ottoman imperial mosque located in Laleli, Fatih, Istanbul, Turkey. The Laleli Mosque was built by Sultan Mustafa III from 1760-1763, designed in the baroque style by Ottoman imperial architect Mehmet Tahir Ağa. The complex was destroyed by a fire in 1783 shortly after its completion and was immediately rebuilt. A fire in 1911 destroyed the madrasah, and subsequent road construction work destroyed many other auxiliary structures to the mosque. The mosque was built on a high terrace over a complex of vaulted shops, whose rents were intended to financially support the **</information>**

<answer> no. <
<warning> My previous action is invalid. If I want to search, I should put the query between **<search>** and **</search>**. If I want to give the final answer, I should put the answer between **<answer>** and **</answer>**. Let me try again. </warning>

<search> Are the Laleli Mosque and Esma Sultan Mansion located in the same neighborhood </search>

<information> Doc 1<## Title: "Esma Sultan Mansion" ##> Esma Sultan Mansion The Esma Sultan Mansion (), a historical yalı () located at Bosphorus in Ortaköy neighborhood of Istanbul, Turkey and named after its original owner Esma Sultan, is used today as a cultural center after being redeveloped. The three-storey brick building was designed by the renowned architect Sarkis Balyan and finished in 1875 next to Ortaköy Mosque. It was presented to Esma Sultan, the daughter of Ottoman Sultan Abdulaziz, as a wedding gift. The mansion remained in the possession of the Ottoman dynasty until 1915. The building was subsequently used first as a tobacco warehouse and then as
Doc 2<## Title: "Esma Sultan Mansion" ##> guests in the ground floor and 600 guests in the first floor. The venue also hosts concerts of the Istanbul International Jazz Festival and the Istanbul International Music Festival. Esma Sultan Mansion The Esma Sultan Mansion (), a historical yalı () located at Bosphorus in Ortaköy neighborhood of Istanbul, Turkey and named after its original owner Esma Sultan, is used today as a cultural center after being redeveloped. The three-storey brick building was designed by the renowned architect Sarkis Balyan and finished in 1875 next to Ortaköy Mosque. It was presented to Esma Sultan, the daughter of Ottoman Sultan Abdulaziz,
Doc 3<## Title: "Laleli Mosque" ##> Laleli Mosque The Laleli Mosque () is an 18th-century Ottoman imperial mosque located in Laleli, Fatih, Istanbul, Turkey. The Laleli Mosque was built by Sultan Mustafa III from 1760–1763, designed in the baroque style by Ottoman imperial architect Mehmet Tahir Ağa. The complex was destroyed by a fire in 1783 shortly after its completion and was immediately rebuilt. A fire in 1911 destroyed the madrasah, and subsequent road construction work destroyed many other auxiliary structures to the mosque. The mosque was built on a high terrace over a complex of vaulted shops, whose rents were intended to financially support the </information>

<answer> no. <
<warning> My previous action is invalid. If I want to search, I should put the query between **<search>** and **</search>**. If I want to give the final answer, I should put the answer between **<answer>** and **</answer>**. Let me try again. </warning>
<warning> Now, I can no longer call **<search></search>** for external information. I must analyze and answer based on the existing information. Now let me make a final thought and give an answer. </warning>

<answer> Yes. </answer>

G.4 SEARCH-R1-V0.3

Search-R1-v0.3 (Incorrect)

Level: Hard **Source:** HotpotQA **Reasoning Type:** Comparison

[Question]

Are the Laleli Mosque and Esma Sultan Mansion located in the same neighborhood?

[Gold Answer]

No

[Titles of Support Evidence Documents (D)]

Laleli Mosque, Esma Sultan Mansion

[Response]

A: <search>are the laleli mosque and esma sultan mansion located in the same neighborhood</search>

<information>Doc 1(Title: ”Esma Sultan Mansion”) Esma Sultan Mansion The Esma Sultan Mansion (), a historical yalı () located at Bosphorus in Ortaköy neighborhood of Istanbul, Turkey and named after its original owner Esma Sultan, is used today as a cultural center after being redeveloped. The three-storey brick building was designed by the renowned architect Sarkis Balyan and finished in 1875 next to Ortaköy Mosque. It was presented to Esma Sultan, the daughter of Ottoman Sultan Abdulaziz, as a wedding gift. The mansion remained in the possession of the Ottoman dynasty until 1915. The building was subsequently used first as a tobacco warehouse and then as
Doc 2(Title: ”Esma Sultan Mansion”) guests in the ground floor and 600 guests in the first floor. The venue also hosts concerts of the Istanbul International Jazz Festival and the Istanbul International Music Festival. Esma Sultan Mansion The Esma Sultan Mansion (), a historical yalı () located at Bosphorus in Ortaköy neighborhood of Istanbul, Turkey and named after its original owner Esma Sultan, is used today as a cultural center after being redeveloped. The three-storey brick building was designed by the renowned architect Sarkis Balyan and finished in 1875 next to Ortaköy Mosque. It was presented to Esma Sultan, the daughter of Ottoman Sultan Abdulaziz,
Doc 3(Title: ”Laleli Mosque”) Laleli Mosque The Laleli Mosque () is an 18th-century Ottoman imperial mosque located in Laleli, Fatih, Istanbul, Turkey. The Laleli Mosque was built by Sultan Mustafa III from 1760-1763, designed in the baroque style by Ottoman imperial architect Mehmet Tahir Ağa. The complex was destroyed by a fire in 1783 shortly after its completion and was immediately rebuilt. A fire in 1911 destroyed the madrasah, and subsequent road construction work destroyed many other auxiliary structures to the mosque. The mosque was built on a high terrace over a complex of vaulted shops, whose rents were intended to financially support the</information>

<answer> Yes </answer>

H FEW-SHOT EXAMPLES

Example 1

Level: Easy **Source:** HotpotQA **Annotation by:** DeepSeek-R1-Distill-Qwen-32B

[Input]

[user]
You are an intelligent reasoning assistant with access to a web search tool to help you obtain accurate information. When answering the user's questions, please adhere to the following guidelines:

1. **Uncertainty Handling:**
   - Encourage the use of web search to solve the problems mentioned during reasoning.
   - If you are uncertain or do not have sufficient knowledge about a specific detail, you **must** perform a web search.
   - To perform a web search the format **must** be in the format mentioned in point 2 Web Search Format
   - You can perform multiple web searches to ensure that the answer is correct.
   - The web search should be initiated only within your internal reasoning (the "thinking" phase), and it should not appear in your final answer to the user.

2. **Web Search Format:**
   - When calling the web search tool, use the following exact format:
     ```
     <begin_search> your query here </end_search>
     ```
   - After you perform a web search, you will receive results enclosed in:
     ```
     <search_result> ... </search_result>
     ```
   - You can then use this information to further refine your reasoning and answer.

3. **Process:**
   - If you encounter any uncertainty or unknown knowledge, embed a web search query within your internal thoughts (surrounded by `<think></think>` tags) using the specified format.
   - Incorporate the search results into your reasoning before finalizing your answer.
   - You should provide your final answer in the format **\boxed**{YOUR_ANSWER}.

Now, please answer the user's question below:
The Voyage that Shook the World was commissioned by what non-profit organisation that has locations in Australia, Canada, New Zealand, South Africa, the United Kingdom, and the United States?

[Response]

<think>
Okay, so I need to figure out which non-profit organization commissioned the voyage known as "The Voyage that Shook the World." The user mentioned that this organization has locations in several countries, including Australia, Canada, New Zealand, South Africa, the UK, and the US. Hmm, that gives me some clues about its global presence.

First, I'm thinking about well-known non-profits that operate internationally. Organizations like the Red Cross come to mind, but I don't recall them being involved in something called "The Voyage that Shook the World." Maybe it's something related to exploration or maritime activities? That could make sense given the word "voyage."

I remember hearing about the HMS Beagle and Charles Darwin's voyage, which was significant in shaping evolutionary theory. Could this be related? There's a project or initiative that recreates historical voyages for educational purposes. I think it's called something like the Beagle Project. Let me check that.

Wait, the Beagle Project is indeed a non-profit organization. They focus on recreating historical scientific voyages, including the one that Darwin took. Their mission is to promote science education and exploration. Given that, it makes sense that they would commission a voyage that has historical significance, hence the name "The Voyage that Shook the World."

To confirm, I should probably look up if the Beagle Project operates in all those countries mentioned. I'll do a quick search to verify their international presence.

[Initiating web search]
<begin_search>Beagle Project locations</end_search>

<search_result> result 1: "Beagle Brigade"
they are placed in adoptive homes. The USDA National Detector Dog Training Center in Newnan, Georgia, receives beagles donated by private owners, breeders, and animal shelters.
result 2: "Beagle station"
Beagle is a station on the Port Authority of Allegheny County's light rail network, located in Bethel Park, Pennsylvania. The street level stop is designed as a small commuter stop, serving area residents who walk to the train so they can be taken toward Downtown Pittsburgh. </search_result>

Hmm, the search results didn't directly mention the Beagle Project having locations in those countries. Maybe I need a different approach. Perhaps I should search specifically for "The Voyage that Shook the World" and see which organization commissioned it.

[Initiating another web search]
<begin_search>The Voyage that Shook the World commissioned by </end_search>

<search_result> result 1: "The Voyage that Shook the World"
The Voyage That Shook The World is a 2009 dramatised documentary film commissioned by Creation Ministries International, a Christian Young Earth creationist organisation, and produced by Fathom Media. It was released to mark the 200th anniversary of Charles Darwin's birth and the 150th anniversary of the

```
    publication of his seminal work ”On the Origin of Species”. A
    historian featured in the film has stated that the creationist
    backing of the film had been concealed when he agreed to take
    part, that the editing of his words could give a false
    impression of his views, and that the film presents a
    historically distorted portrait of Darwin. Creation Ministries
    agreed that they had set up a ”front company” to approach
    experts. They denied any deception and stated that one of the
    interviewees had admitted that while the producers choose
    comments they ”didn't distort what we said”, and compared their
    approach to that used by the BBC in making documentaries. The
    three historians featured in the film subsequently issued a
    statement that they had been misrepresented by the film company'
    s selective reconstruction of Darwin's voyage. CMI countered
    these claims with extended quotes from the interviews of the
    historians. Their response was described as appearing to be
    sound in specific aspects by American skeptic Jim Lippard, who
    had not seen the
result 2: ”Henry Roberts (Royal Navy officer)”
Henry Roberts (1756-1796) was an officer in the Royal Navy who
    served with Captain Cook on his last two voyages. Roberts served
     as lieutenant on Cook's HMS ”Discovery”, where he was entrusted
     with many hydrographic and cartographic tasks, and also met
    then-midshipman George Vancouver. Roberts spent many years after
     that voyage preparing the detailed charts that went into Cook's
     posthumous ”Atlas”. Roberts also served on where he was to make
     a painting of the ship.  In 1790, Roberts was appointed to
    command a newly built HMS ”Discovery” on another round-the-world
     voyage and selected George Vancouver as his first lieutenant.
    However, the Nootka Crisis called both men to duty elsewhere,
    and upon its resolution, Vancouver was given command of the
    historic voyage. Roberts died on 25 August 1796, Captain of HMS
    ”Undaunted” in the West Indies, where he contracted yellow fever.
     Vancouver named Point Roberts after Henry Roberts.
    </search_result>

Ah, from the first result, it seems that ”The Voyage that Shook the
    World” was commissioned by Creation Ministries International (
    CMI). They're a Christian Young Earth creationist organization.
    This aligns with the fact that the voyage in question is likely
    referring to Charles Darwin's journey on the HMS Beagle, which
    had a significant impact on the world through his theories on
    evolution.

CMI operates in multiple countries, including those listed by the
    user: Australia, Canada, New Zealand, South Africa, the UK, and
    the US. This fits perfectly with the question's details. So,
    despite my initial thought about the Beagle Project, the correct
     answer is Creation Ministries International.
</think>

The non-profit organization that commissioned ”The Voyage that Shook
     the World” is Creation Ministries International (CMI).

\boxed{Creation Ministries International}
```

Example 2

Level: Easy **Source:** HotpotQA **Annotation by:** DeepSeek-R1-Distill-Qwen-32B

[Input]

[user]
You are an intelligent reasoning assistant with access to a web search tool to help you obtain accurate information. When answering the user's questions, please adhere to the following guidelines:

1. **Uncertainty Handling:**
   - Encourage the use of web search to solve the problems mentioned during reasoning.
   - If you are uncertain or do not have sufficient knowledge about a specific detail, you **must** perform a web search.
   - To perform a web search the format **must** be in the format mentioned in point 2 Web Search Format
   - You can perform multiple web searches to ensure that the answer is correct.
   - The web search should be initiated only within your internal reasoning (the "thinking" phase), and it should not appear in your final answer to the user.

2. **Web Search Format:**
   - When calling the web search tool, use the following exact format:
     ```
     <begin_search> your query here </end_search>
     ```
   - After you perform a web search, you will receive results enclosed in:
     ```
     <search_result> ... </search_result>
     ```
   - You can then use this information to further refine your reasoning and answer.

3. **Process:**
   - If you encounter any uncertainty or unknown knowledge, embed a web search query within your internal thoughts (surrounded by `<think></think>` tags) using the specified format.
   - Incorporate the search results into your reasoning before finalizing your answer.
   - You should provide your final answer in the format **\boxed**{YOUR_ANSWER}.

Now, please answer the user's question below:
The World Classical Network (WCN) is an internet-delivery classical radio broadcasting service owned by which company, in Hyannis, the largest of the seven villages in the town of Barnstable, Massachusetts, in the United States?

[Response]

<think>
Okay, so I need to figure out which company owns the World Classical Network (WCN) based in Hyannis, Massachusetts. Hmm, I don't know much about WCN, so I'll have to look it up. Let me think about how to approach this.

First, I know that Hyannis is part of Barnstable, Massachusetts. It's a town with several villages, and Hyannis is the largest one.

So, WCN is located there. But I'm not sure which company operates it. Maybe it's a local media company or part of a larger network.

I remember that sometimes classical music stations are operated by public radio networks or independent broadcasters. Maybe WCN is part of a bigger organization. Alternatively, it could be a subsidiary of a media conglomerate.

To get accurate information, I should perform a web search. I'll search for "World Classical Network owner Hyannis Massachusetts" to see if any reliable sources mention the company. Let me structure the search query properly.

<begin_search>World Classical Network owner Hyannis Massachusetts</end_search>

<search_result> result 1: "World Classical Network"
The World Classical Network (WCN) is an internet-delivery classical radio broadcasting service owned by CCB Media in Hyannis, Massachusetts. The service has been in existence since February 1998, when it was owned and operated by Charles River Broadcasting. It filled a void left by the demise of the U. S. version of Classic FM, the classical music satellite service of SW Networks. Stations. WCN programming is carried on Cape Cod Broadcasting's WFCC-FM, licensed to Chatham, Massachusetts. WFCC is the WCN flagship station. The service is also syndicated, with the following affiliates: Hosts. Current hosts on the World Classical Network include Mark Calder, Don Spencer, and Larry King. Longtime classical announcer Steve Murphy joined the WCN as morning host on September 16, 2008.
result 2: "Charles River Broadcasting"
Charles River Broadcasting was the owner of three classical music stations, one classic rock station, one CNN Headline News relay in Rhode Island, and a syndicated classical satellite delivery program service. History. The company started in 1948 when WCRB first signed on at 1330 AM. In 1954, Charles River Broadcasting added WCRB-FM at 102.5 FM. In the mid 1970s, WCRB's programming was removed from the 1330 AM signal, which was relaunched as WHET, with a big-band/adult standards format. (WHET was sold in 1978 and is now WRCA.) By the early 1990s, Charles River Broadcasting and WCRB entered into a commitment for the station to continue running classical music until 2092. Later in the decade, the company purchased WFCC-FM Chatham, WKPE-FM Orleans (both on Cape Cod), WCRI Block Island/Westerly (formerly WVBI), and WCNX Hope Valley, Rhode Island (formerly WJJF) (both in Rhode Island) from their respective local owners. In addition, a satellite service named the World Classical Network was started by Charles River Broadcasting. The company explored a sale possibility for </search_result>

Alright, looking at the search results, the first result mentions that WCN is owned by CCB Media in Hyannis, Massachusetts. It also notes that Charles River Broadcasting was the original owner but later became CCB Media. The second result confirms that Charles River Broadcasting started the World Classical Network and later sold it or rebranded under CCB Media.

So, putting this together, the current owner of WCN is CCB Media. They took over from Charles River Broadcasting, which had initially established the service. Therefore, the company in Hyannis that owns WCN is CCB Media.

```
</think>

The World Classical Network (WCN) is owned by CCB Media in Hyannis,
    Massachusetts.

\boxed{CCB Media}
```

Example 3

Level: Medium **Source:** HotpotQA **Annotation by:** DeepSeek-R1-Distill-Qwen-32B

[Input]

```
[user]
You are an intelligent reasoning assistant with access to a web
    search tool to help you obtain accurate information. When
    answering the user's questions, please adhere to the following
    guidelines:

1. Uncertainty Handling:
   - Encourage the use of web search to solve the problems mentioned
     during reasoning.
   - If you are uncertain or do not have sufficient knowledge about
    a specific detail, you must perform a web search.
   - To perform a web search the format must be in the format
    mentioned in point 2 Web Search Format
   - You can perform multiple web searches to ensure that the answer
     is correct.
   - The web search should be initiated only within your internal
    reasoning (the "thinking" phase), and it should not appear in
    your final answer to the user.

2. Web Search Format:
   - When calling the web search tool, use the following exact
    format:
     ```
     <begin_search> your query here </end_search>
     ```
   - After you perform a web search, you will receive results
    enclosed in:
     ```
     <search_result> ... </search_result>
     ```
   - You can then use this information to further refine your
    reasoning and answer.

3. Process:
   - If you encounter any uncertainty or unknown knowledge, embed a
    web search query within your internal thoughts (surrounded by `
    <think></think>` tags) using the specified format.
   - Incorporate the search results into your reasoning before
    finalizing your answer.
   - You should provide your final answer in the format \boxed{
    YOUR_ANSWER}.

Now, please answer the user's question below:
The seventh FIFA Women's World Cup, in which Sukanya Chor
    Charoenying played, occurred in what year?
```

[Response]

<think>
Alright, I need to figure out the year when the seventh FIFA Women's World Cup took place, specifically the one in which Sukanya Chor Charoenying played. Let me start by recalling what I know about the FIFA Women's World Cup.

I remember that the first Women's World Cup was held in 1991, so each subsequent tournament would follow every four years. Let me count them out:

1. 1991 - First edition
2. 1995 - Second
3. 1999 - Third
4. 2003 - Fourth
5. 2007 - Fifth
6. 2011 - Sixth
7. 2015 - Seventh

Wait, does that add up correctly? From 1991 to 1995 is four years, then each subsequent is another four. So yes, the seventh would be 2015. But I'm not entirely sure, so I should double-check this information.

Additionally, I need to confirm whether Sukanya Chor Charoenying participated in the 2015 tournament. I'm not certain about her participation in that specific year, so I should look into her career timeline as well.

Let me structure my thoughts:

1. Confirm the year of the seventh FIFA Women's World Cup.
2. Verify if Sukanya Chor Charoenying played in that edition.

Starting with the first point, I think the seventh edition was in 2015, but I want to make sure. I'll perform a web search to confirm the years of each Women's World Cup.

<begin_search>FIFA Women's World Cup editions years</end_search>

<search_result> result 1: "FIFA Women's World Cup"
The FIFA Women's World Cup is an international football competition contested by the senior women's national teams of the members of Fédération Internationale de Football Association (FIFA), the sport's international governing body. The competition has been held every four years since 1991, when the inaugural tournament, then called the FIFA Women's World Championship, was held in China. Under the tournament's current format, national teams vie for 23 slots in a three-year qualification phase. The host nation's team is automatically entered as the 24th slot. The tournament proper, alternatively called the "World Cup Finals", is contested at venues within the host nation(s) over a period of about one month. The eight FIFA Women's World Cup tournaments have been won by four national teams. The United States has won four times, and is the current champion after winning it at the 2019 tournament in France. The other winners are Germany, with two titles, and Japan and Norway with one title each. Six countries have hosted the Women's World Cup. China and the United States have each hosted the tournament twice, while Canada, France, Germany, and Sweden have each hosted it once. Format. Qualification. Qualifying tournaments are held within

the six FIFA continental zones (Africa, Asia, North and Central America and Caribbean, South
result 2: ”FIFA Women's World Cup”
automatically chosen to host the 2007 FIFA Women's World Cup. Germany hosted the 2011 FIFA Women's World Cup, as decided by vote in October 2007. In March 2011, FIFA awarded Canada the right to host the 2015 FIFA Women's World Cup. The 2015 edition saw the field expand from 16 to 24 teams. During the 2015 FIFA Women's World Cup, both Formiga of Brazil and Homare Sawa of Japan appeared in their record sixth World Cup, a feat that had never been achieved before by either female or male players. Christie Pearce is the oldest player to ever play in a Women's World Cup match, at the age of 40 years. In March 2015, FIFA awarded France the right to host the 2019 FIFA Women's World Cup over South Korea. Trophy. The current trophy was designed in 1998 for the 1999 tournament, and takes the form of a spiral band, enclosing a football at the top, that aims to capture the athleticism, dynamism, and elegance of international women's football. In the 2010s, it was fitted with a cone-shaped base. Underneath the base, the name of each of the tournament's previous winners is engraved. The trophy **</search_result>**

From the search results, I see that the 2015 FIFA Women's World Cup was the seventh edition. It was hosted by Canada, and it was the first time the tournament expanded to 24 teams. So that confirms the year as 2015.

Now, moving on to the second point: Sukanya Chor Charoenying's participation. I need to check if she played in the 2015 Women's World Cup.

<begin_search>Sukanya Chor Charoenying FIFA Women's World Cup 2015 **</end_search>**

<search_result> result 1: ”Sukanya Chor Charoenying”
Sukanya Chor Charoenying (born 24 November 1987) is a Thai women's international footballer who plays as a goalkeeper. She is a member of the Thailand women's national football team. She was part of the team at the 2015 FIFA Women's World Cup. On club level she plays for Air Force United in Thailand.
result 2: ”Sudarat Chuchuen”
Sudarat Chuchuen (; born 19 June 1997) is a Thai footballer who plays as a defender for Sisaket FC and the Thailand women's national team. She was selected for the 2019 FIFA Women's World Cup. **</search_result>**

The first result clearly states that Sukanya Chor Charoenying was part of the Thailand women's national team at the 2015 FIFA Women's World Cup. The second result mentions another player, Sudarat Chuchuen, who played in the 2019 edition, but that's not relevant here.

Therefore, combining both pieces of information, the seventh FIFA Women's World Cup occurred in 2015, and Sukanya Chor Charoenying was indeed a participant in that tournament.
</think>

The seventh FIFA Women's World Cup took place in 2015, and Sukanya Chor Charoenying was part of the Thailand team during that event.

\boxed{2015}

Example 4

Level: Easy **Source:** HotpotQA **Annotation by:** DeepSeek-R1-Distill-Qwen-32B

[Input]

```
[user]
You are an intelligent reasoning assistant with access to a web
    search tool to help you obtain accurate information. When
    answering the user's questions, please adhere to the following
    guidelines:

1. Uncertainty Handling:
   - Encourage the use of web search to solve the problems mentioned
     during reasoning.
   - If you are uncertain or do not have sufficient knowledge about
    a specific detail, you must perform a web search.
   - To perform a web search the format must be in the format
    mentioned in point 2 Web Search Format
   - You can perform multiple web searches to ensure that the answer
     is correct.
   - The web search should be initiated only within your internal
    reasoning (the "thinking" phase), and it should not appear in
    your final answer to the user.

2. Web Search Format:
   - When calling the web search tool, use the following exact
    format:
     ```
     <begin_search> your query here </end_search>
     ```
   - After you perform a web search, you will receive results
    enclosed in:
     ```
     <search_result> ... </search_result>
     ```
   - You can then use this information to further refine your
    reasoning and answer.

3. Process:
   - If you encounter any uncertainty or unknown knowledge, embed a
    web search query within your internal thoughts (surrounded by `
    <think></think>` tags) using the specified format.
   - Incorporate the search results into your reasoning before
    finalizing your answer.
   - You should provide your final answer in the format \boxed{
    YOUR_ANSWER}.

Now, please answer the user's question below:
Who founded the Honest Conpnay along with Brian Lee
```

[Response]

```
<think>
Okay, so I need to figure out who founded the Honest Company along
    with Brian Lee. I remember hearing about the Honest Company
    being a company that makes baby products and other household
    items, but I'm not exactly sure about their founders. I think it
```

```
 might involve someone well-known, maybe from the entertainment
industry?

Wait, wasn't there a famous actor involved? I think it was someone
like Jessica Alba. I recall seeing her name associated with
Honest Company somewhere. Let me try to remember. Yes, Jessica
Alba is an actress, and she started a company focused on natural
 and safe products for babies. That sounds right.

But I'm not entirely certain if Brian Lee is her co-founder or if
there's another person. Maybe Brian Lee is another partner or
investor? I'm a bit fuzzy on the details. To make sure, I should
 look up the founders of the Honest Company.

I'll perform a quick search to confirm.

<begin_search>founders of the honest company</end_search>

<search_result> result 1: "The Honest Company"
The Honest Company is an American consumer goods company, founded by
 actress Jessica Alba. It emphasizes household products to
supply the marketplace for ethical consumerism. The company had
$250 million in 2016 sales, and was valued shy of $1 billion as
of October 2017. The Honest Company has raised multiple rounds
of venture capital, and was anticipating an initial public
offering as of 2016. Honest serves the United States and Canada,
 and planned to launch its beauty line products in Western
Europe in 2019. History. Company founder Jessica Alba was
inspired by the 2008 birth of her first child, Honor, and her
own history of childhood illnesses to create a company that
provided an alternative to the prevalent baby products with
ingredients such as petrochemicals and synthetic fragrances. She
 was compelled to become serious about this venture when one of
her mother's baby laundry recommendations caused her to have a
welt outbreak. It took Alba three years to find her business
partners Brian Lee, Sean Kane, and Christopher Gavigan. Despite
advice that she should start small with a singular focus, Alba
launched the company in 2012 with 17 products. Christopher
Gavigan wrote the book, Healthy Child Healthy World, that
addresses all the different toxins that children are exposed to
in traditional household
result 2: "Honest Tea"
Honest Tea (U.S.) is a bottled organic tea company based in Bethesda
, Maryland. It was founded in 1998 by Seth Goldman and Barry
Nalebuff. The company is a wholly owned subsidiary of The Coca-
Cola Company. The name is a pun on the word "honesty". History.
Honest Tea was founded in 1998 by Seth Goldman, a graduate of
Harvard and the Yale School of Management who got the idea to
found a beverage company while he was at business school. There
were two key ideas behind the company: to bring beverages to
market that were organic and not as heavily sweetened, and to
conduct the business using fair trade principles. Goldman and
one of his Yale professors, Barry Nalebuff, conceived the idea
for the company and raised $500,000 in startup funds from
friends and family. The first order for Honest Tea was from
Fresh Fields, an East Coast natural grocery chain that was later
 acquired by Whole Foods Market. By 2006, Honest Tea had revenue
 of $13.5 million and was selling about 1.5 million cases a year.
 The Coca-Cola Company purchased a 40% stake of the U.S. company
 at $43 million in 2008. It bought the rest of the company in
</search_result>
```

Alright, from the search results, it seems that the Honest Company, which is different from Honest Tea, was indeed founded by Jessica Alba. The other co-founders listed are Brian Lee, Sean Kane, and Christopher Gavigan. So, Brian Lee is definitely one of the co-founders alongside Jessica Alba.

Just to clarify, the search result mentions that Jessica Alba found the company with Brian Lee, Sean Kane, and Christopher Gavigan. They started it in 2012 with 17 products. So, the answer should be that Jessica Alba founded the Honest Company along with Brian Lee, Sean Kane, and Christopher Gavigan. However, the question specifically asks who founded it along with Brian Lee, so the primary co-founder is Jessica Alba.
</think>

The Honest Company was founded by Jessica Alba along with Brian Lee, Sean Kane, and Christopher Gavigan. Therefore, the co-founder alongside Brian Lee is Jessica Alba.

\boxed{**Jessica Alba**}