

UR²: UNIFY RAG AND REASONING THROUGH REINFORCEMENT LEARNING

Anonymous authors

Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) have shown remarkable capabilities through two complementary paradigms: Retrieval-Augmented Generation (RAG), which enhances knowledge grounding, and Reinforcement Learning from Verifiable Rewards (RLVR), which optimizes complex reasoning abilities. However, these two capabilities are often developed in isolation, and existing efforts to unify them remain narrow in scope—typically limited to open-domain QA with fixed retrieval settings and task-specific constraints. This lack of integration constrains generalization and limits the applicability of RAG-RL methods to broader domains. To bridge this gap, we propose **UR²** (Unified **R**AG and **R**easoning), a general framework that unifies retrieval and reasoning through reinforcement learning. UR² introduces two key contributions: a difficulty-aware curriculum training that selectively invokes retrieval only for challenging problems, and a hybrid knowledge access strategy combining domain-specific offline corpora with LLM-generated summaries. These components are designed to enable dynamic coordination between retrieval and reasoning, improving adaptability across a diverse range of tasks. Experiments across open-domain QA, MMLU-Pro, medical, and mathematical reasoning tasks demonstrate that UR² (built on Qwen-2.5-3/7B and LLaMA-3.1-8B) significantly outperforms existing RAG and RL methods, achieving comparable performance to GPT-4o-mini and GPT-4.1-mini on several benchmarks. We will release all code, models, and data upon submission.

1 INTRODUCTION

Large Language Models (LLMs) have achieved remarkable performance across diverse tasks by incorporating external knowledge (Retrieval-Augmented Generation, RAG) (Lewis et al., 2020; Borgeaud et al., 2022; Izacard et al., 2022) and optimizing reasoning through reinforcement learning with verifiable rewards (RLVR) (Guo et al., 2025a). RAG methods enable LLMs to access external knowledge, while RLVR shows strong gains on mathematical and logical reasoning (Zeng et al., 2025; Chen et al., 2025). Motivated by these successes, recent work has begun to integrate retrieval and reasoning: for example, Search-o1 (Li et al., 2025) embeds an agentic RAG workflow into the LLM’s chain-of-thought, and RAG-Gym (Xiong et al., 2025) proposes a unified RL-based training framework for RAG agents. Similarly, RAG-RL methods—which learn to invoke retrieval through RL—such as R1-Searcher (Song et al., 2025a) and Search-R1 (Jin et al., 2025) use RLVR to train models on *when* and *what* to retrieve during reasoning, improving performance in open-domain QA.

Despite recent progress, RAG-RL frameworks remain limited in scope. Most methods focus narrowly on open-domain QA, with retrieval tied to fixed reasoning steps or static knowledge sources like Wikipedia. However, paradigms that work well on open-domain QA often fail to transfer to broader domains. **Two key limitations persist:** (1) models struggle to achieve optimal reasoning-retrieval trade-offs, often **over-emphasizing one component**; (2) retrieved documents **contain significant noise**, and current prompt-only approaches lack effective noise handling mechanisms, potentially degrading to basic Chain-of-Thought reasoning. For instance, R1-Searcher and Search-R1 assume access to Wikipedia, ill-suited for tasks requiring specialized or real-time information. While methods like DeepResearcher attempt training in real web environments, they face inefficiencies due to the noisy and unstructured nature of online data (Zheng et al., 2025). Other methods like ZeroSearch (Sun et al., 2025), use LLM-generated corpora to simulate retrieval, avoiding API costs but risking hallucination and loss of real-world complexity.

To address the limitations of existing RAG-RL approaches—such as static retrieval, limited domain generalization, and poor robustness in noisy environments—we propose a general and adaptive framework, **UR²** (**U**nified **R**AG and **R**easoning), which uses RL to dynamically coordinate retrieval and reasoning. Unlike prior methods that rely solely on static corpora (e.g., Wikipedia) or simulate retrieval with synthetic content, UR² combines both: it leverages task-specific offline corpora for accurate grounding, augmented with LLM-generated summaries for efficiency and generalization. To address the imbalance between retrieval and reasoning in prior methods, we design a difficulty-aware curriculum that adaptively controls when to trigger retrieval during training. Specifically, retrieval is used only for hard instances, encouraging the model to rely on internal reasoning when possible and to learn retrieval strategies only when necessary. This reduces retrieval overhead, improves query quality on challenging questions, and preserves reasoning capability across tasks.

We train UR² on Qwen-2.5-3B/7B-Instruct (Yang et al., 2024) and LLaMA-3.1-8B-Instruct (Dubey et al., 2024) across MMLU-Pro, Medicine, Math, and open-domain QA. During training, these models spontaneously develop key cognitive behaviors: self-verification through retrieval, intermediate reasoning validation, and hypothesis revision based on external evidence. UR² outperforms previous state-of-the-art (SOTA) methods by **5.8%** (7B) and **19.0%** (3B) on average, with peak gains of **9.5%** and **29.6%**. Notably, our 7B model matches GPT-4o-mini and GPT-4.1-mini¹, and generalizes well across domains and model architectures.

Our main contributions are summarized as follows:

- We propose the first unified retrieval-reasoning RL framework that adapts to diverse tasks beyond open-domain QA, representing an important milestone for AI systems combining parametric and external knowledge.
- We design a unified data representation and training scheme bridging retrieval and reasoning, with difficulty-aware curricula and LLM-summarized corpora for accurate grounding and broad generalization, implemented via a modular two-stage framework.
- Comprehensive experiments demonstrate that UR² surpasses advanced RAG and RL methods without expert demonstrations and generalizes robustly across domains.

2 RELATED WORK

2.1 RETRIEVAL-AUGMENTED GENERATION

RAG enhances LLMs by incorporating external information to reduce hallucinations (Gao et al., 2023). Early RAG methods concatenate retrieved documents with input prompts (Lewis et al., 2020; Izacard et al., 2022; Borgeaud et al., 2022). Subsequent approaches have evolved in multiple directions: advanced RAG methods incorporate sophisticated retrieval and re-ranking strategies (Gao et al., 2023; Peng et al., 2024); post-hoc verification methods address hallucinations by retrieving documents based on generated responses (Li et al., 2024; Sun et al., 2024); and Graph-based RAG methods integrate knowledge graphs for multi-hop reasoning (Edge et al., 2024; Hu et al., 2025b; Peng et al., 2024). Recent RAG-RL frameworks have explored retrieval integration during training via real-time or synthetic retrieval (Zheng et al., 2025; Sun et al., 2025). However, these approaches remain constrained by static retrieval strategies, limited domain generalization, and **inability to dynamically coordinate retrieval with reasoning** across diverse task types.

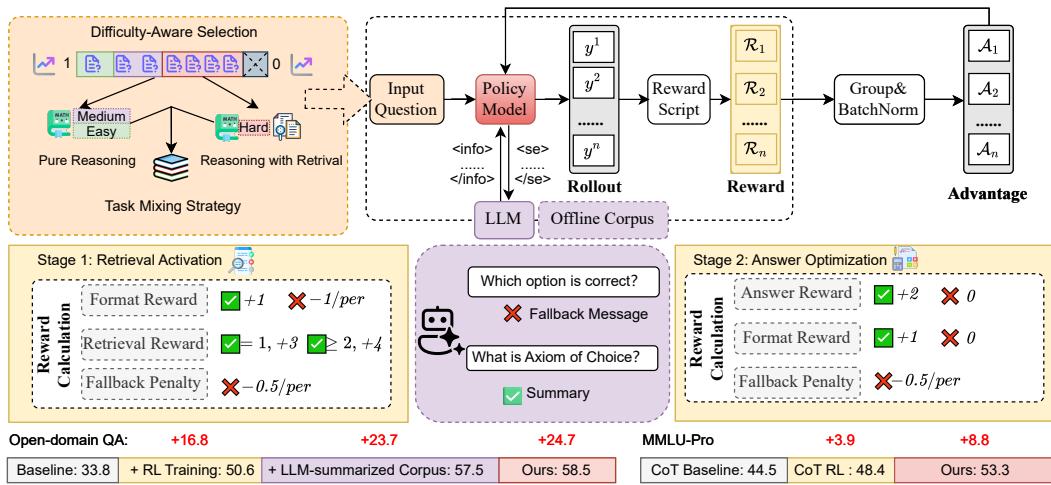
2.2 REINFORCEMENT LEARNING FOR RETRIEVAL-ENHANCED REASONING

RL has emerged as a key technique for significantly improving LLM capabilities, evolving from early policy gradient methods such as REINFORCE (Williams, 1992) to more advanced algorithms like PPO (Schulman et al., 2017) and GRPO (Shao et al., 2024). Recent methods, including ARENA (Ren et al., 2025), Search-R1 (Jin et al., 2025), and R1-Searcher (Song et al., 2025a), demonstrate that RL enables LLMs to effectively learn multi-step reasoning and retrieval strategies without requiring human feedback. These works collectively highlight a clear shift from fixed retrieval heuristics to learned, RL-driven retrieval policies, which form the foundation for our unified framework, with retrieval becoming **increasingly parameterized** rather than merely prompt-guided.

¹<https://chat.openai.com/>

108 3 METHOD
109

110 We propose UR², a general framework that tightly integrates retrieval-based grounding with explicit
111 step-by-step reasoning via RL. Unlike previous approaches restricted to open-domain QA or reliant
112 upon static corpora, UR² supports a broad range of tasks, including mathematical problem solving
113 and domain-specific QA. To achieve this versatility, UR² leverages a LLM-summarized retrieval
114 corpus (Section 3.1.1) and a difficulty-aware curriculum that adapts training based on task hardness
115 and knowledge demands (Section 3.1.2).



133 Figure 1: Overview of the UR² training pipeline. The top illustrates LLM-summarized retrieval
134 corpus, difficulty-aware curriculum design and a two-stage reward design for retrieval activation and
135 answer optimization. The bottom horizontal bars indicate: (1) Open-domain QA: ablation results
136 under LLM-as-a-judge; and (2) MMLU-Pro: comparison with baselines using EM score.

137
138 3.1 DIFFICULTY-AWARE TRAINING WITH HYBRID KNOWLEDGE ACCESS
139

140 3.1.1 LLM-SUMMARIZED RETRIEVAL CORPUS
141

142 UR² employs a LLM-summarized retrieval corpus designed to accommodate diverse task domains,
143 comprising:

- 144
- 145 • Domain-specific offline corpora (e.g., curated medical knowledge bases, full Wikipedia
146 content, or Wikipedia abstracts);
 - 147 • Concise summary or **fallback response** generated by LLMs, following the structured ap-
148 proach of Search-01 (Li et al., 2025), which enables the system to reject queries requiring
149 complex reasoning beyond the scope of retrieval (see Appendix D.3).

151 This hybrid corpus design enhances retrieval accuracy, reduces hallucinations, and improves gener-
152 alization across a variety of reasoning scenarios.

153
154 3.1.2 DIFFICULTY-AWARE CURRICULUM DESIGN
155

156 We organize this part into two components: (1) training data selection based on difficulty levels; and
157 (2) task mixing strategy to balance retrieval and reasoning exposure.

158 **Trainning Data Selection.** To promote fine-grained reasoning and retrieval behaviors, we cate-
159 gorize training samples by their difficulty levels. For each question, we perform 20 rollouts using
160 Qwen-2.5-7B-Instruct and compute the average performance score (s). Based on the score s , ques-
161 tions are categorized into three difficulty levels: Easy ($0.8 \leq s \leq 1.0$), Medium ($0.5 \leq s < 0.8$),
162 and Hard ($0.2 \leq s < 0.5$).

Following prior studies (Yu et al., 2025; Guo et al., 2025b), instances with extremely low performance scores ($s < 0.2$) are filtered, as overly difficult samples hinder effective learning. We adopt a sampling ratio of 7:2:1 for hard, medium, and easy questions, prioritizing challenging examples to enhance reasoning and retrieval capabilities.

Task Mixing Strategy. To effectively balance retrieval and reasoning capabilities, we design two strategic task mixtures:

- *Mathematical reasoning with open-domain QA*: Mathematical tasks are separated by difficulty. Hard mathematical problems use retrieval-augmented prompting (Figure 2), while the others rely on pure step-by-step reasoning. QA data consistently uses retrieval since these questions need external knowledge.
- *Multiple-choice reasoning tasks*: We combine MedQA training data and synthetic MMLU-style datasets. Among these, most hard-level questions are used for retrieval-augmented training with a smaller portion for direct reasoning, maintaining an overall 1:1 ratio between the two approaches.

This controlled task composition brings several benefits. First, it ensures diverse exposure to both retrieval-intensive and reasoning-intensive formats, helping the model generalize across different tasks. Second, by associating retrieval usage with question difficulty, the model learns to rely on external knowledge only when necessary, rather than overusing retrieval indiscriminately. Third, this approach saves computational resources by activating retrieval only when needed for hard problems, preserving direct reasoning for simple cases. More detailed experimental configurations are provided in the Section 4 and Appendix C.1.

3.2 TWO-STAGE OPTIMIZATION FOR UR²

Given the limited tool invocation capabilities of base models, especially in reasoning-integrated scenarios, we design a two-stage optimization framework to systematically develop retrieval skills and reasoning proficiency. We train UR² using REINFORCE++ (Guo et al., 2025a), a streamlined variant of PPO tailored. To prevent overfitting to retrieved content, we adopt retrieval masking (Song et al., 2025a; Jin et al., 2025). Our implementation is based on the REINFORCE++-baseline provided by OpenRLHF (Hu et al., 2024).

The training objective is defined as:

$$J_{\text{UR}^2}(\theta) = \mathbb{E}_{x, \{y^i\}} \left[\frac{1}{G} \sum_{i=1}^G \frac{1}{|y^i|} \sum_{t=1}^{|y^i|} y_t^i \cdot r_{i,t} \cdot \hat{A}_{i,t} \right] \quad (1)$$

where the importance weight is:

$$r_{i,t} = \frac{\pi_\theta(y_t^i | x, y_{<t}^i; o_i)}{\pi_{\text{old}}(y_t^i | x, y_{<t}^i; o_i)} \quad (2)$$

and the normalized advantage is:

$$\hat{A}_{i,t} = \text{Norm}_{\text{batch}}(\text{Norm}_{\text{group}}(R_i - b)) \quad (3)$$

The advantage $\hat{A}_{i,t}$ is computed by subtracting the group-level reward baseline and applying normalization across the group and batch to improve learning stability. Here, x denotes the input prompt, $\{y^i\}$ are the sampled trajectories, o_i is the retrieved context and b is the group-level baseline (mean of R_i). See Appendix C.1 and Section 4.4 for detailed implementation.

3.2.1 RAG-BASED ROLLOUT

UR² enables the model to issue retrieval queries during reasoning rather than pre-retrieving all information upfront. As illustrated in Figure 2, the prompting mechanism enforces key principles: queries target single facts grounded in external knowledge, retrieval occurs when needed during the reasoning process, and strict format constraints using special tokens demarcate retrieval actions.

This design allows the model to strategically leverage external knowledge by learning *when* to retrieve and *what* to query for purposeful and grounded reasoning.

You are solving a multiple-choice question. Analyze each option step by step and select the best choice. If you're uncertain about any fact, you may issue a **search query** like this: `<se>` a concise query (under 20 words) `</se>`

- You may issue **multiple queries** during your reasoning.
- Each query should focus on **only one specific fact or concept** and **Avoid combining multiple facts in a single query**.

[Examples omitted here]

- You may use **up to four queries** in total — use them wisely.

When documents are returned in the format: `<info>` ... (search results here) `</info>`, integrate the retrieved information into your reasoning to refine your analysis and reach a well-supported conclusion.

Finally, give your answer in this format: the correct answer is: A, B, C, D, etc.

Figure 2: Instruction prompt used to guide retrieval-augmented reasoning in UR². See Appendix D.2 for details.

3.2.2 STAGE 1: RETRIEVAL CAPABILITY ACTIVATION

We use UR² with Qwen-2.5-7B-Instruct on mathematical and open-domain QA tasks as an example. In Stage 1, the model trains on mathematical problems requiring retrieval calls in the specified format (Figure 2). The objective is not answer accuracy, but to enforce correct usage of the retrieval tool and promote retrieval-invoking behavior. This specialized training runs for only 10 steps. Further details on task setup and extensions to other models are provided in Appendix C.5.

The total reward is:

$$R_{i,\text{stage1}} = R_{i,\text{format}} + R_{i,\text{retrieval}} - P_{i,\text{fallback}} \quad (4)$$

where (1) **Format Reward**: +1 for fully compliant output; -1 per violation (e.g., malformed tags, overlength queries, missing retrieval, or illegal tokens); (2) **Retrieval Reward**: +3 for one valid query, +4 for two or more; (3) **Fallback Penalty**: -0.5 per fallback fault.

This stage equips the model with retrieval capabilities and promotes effective integration of retrieved information during generation.

3.2.3 STAGE 2: ANSWER QUALITY OPTIMIZATION

Building on Stage 1, we incorporate correctness feedback to refine generation quality while preserving retrieval behaviors. The updated reward function is:

$$R_{i,\text{stage2}} = R_{i,\text{answer}} + R_{i,\text{format}} - P_{i,\text{fallback}} \quad (5)$$

where (1) **Answer Reward**: +2 for correct answers, 0 for incorrect; (2) **Format Reward**: +1 for fully valid format; 0 otherwise; (3) **Fallback Penalty**: -0.5 per fallback fault.

By decoupling retrieval skill acquisition (Stage 1) from generation optimization (Stage 2), we ensure stable convergence and interpretable credit assignment across complex reasoning trajectories.

4 EXPERIMENTAL SETTINGS

4.1 TRAINING DATASETS

We build a unified training set covering math (SimpleZoo-RL (Zeng et al., 2025)), open-domain QA (R1-Searcher (Song et al., 2025a)), and multi-choice medical QA (MedQA (Jin et al., 2021)). For MMLU-Pro(Wang et al., 2024a) domains (philosophy, history, economics), we generate synthetic questions via Qwen-3-32B². After deduplication and data selection using Qwen-2.5-7B-Instruct on 20 rollouts per question, we obtain 3K samples each for math, open-domain QA, and MedQA, and 2K samples for each MMLU-Pro domain. Details are in Appendix C.1.

²<https://huggingface.co/Qwen/Qwen-3-32B>

270 4.2 EVALUATION BENCHMARKS
271

272 We evaluate generalization across four task families: (1) **Math Reasoning**: MATH500
 273 (in-domain) (Hendrycks et al., 2021), Minerva (OOD) (Lewkowycz et al., 2022); metric:
 274 LLM-as-a-judge. (2) **Medical QA**: MedQA (5-choice, in) (Jin et al., 2021), MMLU-Pro Medical
 275 (M-Med, OOD); metric: EM. (3) **MMLU-Pro**: Philosophy, History, Economics (in), Law
 276 (OOD); metric: EM. (4) **Open-Domain QA**: HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (Ho
 277 et al., 2020) (in); Bamboogle (Press et al., 2023), MusiQue (Trivedi et al., 2022) (OOD); metrics:
 278 F1 and LLM-as-a-judge.
 279

280 4.3 BASELINES
281

282 We compare UR² to: (1) **Vanilla Methods**: Chain-of-Thought (Kojima et al., 2022), Standard
 283 RAG (Borgeaud et al., 2022; Izacard et al., 2022) (top- k =10). (2) **Advanced RAG Methods**:
 284 Search-o1 (Li et al., 2025), Self-Ask (Press et al., 2023), and RAT (Wang et al., 2024b), which
 285 combine reasoning with retrieval using prompt. (3) **CoT-RL Methods**: R1-like methods includ-
 286 ing Open-Reasoner-Zero (Hu et al., 2025a), SimpleRL-Zoo (Zeng et al., 2025), and General-
 287 Reasoner (Ma et al., 2025). (4) **RAG-RL Methods**: R1-Searcher (Song et al., 2025a), R1-
 288 Searcher++ (Song et al., 2025b), Search-R1 (Jin et al., 2025), and ZeroSearch (Sun et al., 2025).
 289 (5) **Vanilla RL**: Baseline implementation following the same training setup and datasets as UR²,
 290 with RAG-RL applied to open-domain QA and CoT-RL to mathematical and multiple-choice tasks.
 291

292 We use Qwen-2.5-3B-Instruct, Qwen-2.5-7B-Instruct, LLaMA-3.1-8B-Instruct, GPT-4o-mini, and
 293 GPT-4.1-mini as backbones (see Appendix C.2 for configs).
 294

295 4.4 IMPLEMENTATION DETAILS
296

297 Retrieval uses BGE-large-en-v1.5³ and the KILT (Petroni et al., 2021) Wikipedia corpus (100-
 298 word segments, 29M documents) following (Song et al., 2025a). Open-domain QA uses Wikipedia
 299 abstract corpus⁴. Unless otherwise noted, all models use GPT-4.1-mini as the summarizer during
 300 training and GPT-4.1 during evaluation with top- k = 10, while mathematical tasks are summarized
 301 by Qwen-3-32B. For evaluation, we sample 500 instances from each benchmark. We use G = 16
 302 rollouts. 7B and 8B models use training batch size 256, rollout batch size 64; 3B doubles both.
 303 Learning rate = 2e-6. Up to 4 retrieval turns are allowed. All models are trained for up to 2 epochs
 304 on 8xA100 GPUs. See Appendix C.1 and C.2 for details.
 305

306 5 EXPERIMENTAL RESULTS
307

308 Our UR² framework achieves SoTA performance across reasoning and retrieval tasks, enabling 7B
 309 models to match or exceed the GPT model family while significantly outperforming existing RAG
 310 and RL-based methods. More comprehensive baseline results can be found in Appendix B.1.
 311

312 5.1 MAIN RESULTS ON REASONING TASKS
313

314 As shown in Table 1, UR² demonstrates substantial improvements across all reasoning tasks on
 315 the Qwen-2.5-7B model, achieving average scores of 53.3% on MMLU-Pro, 65.9% on Medicine,
 316 and 71.0% on Math benchmarks, representing gains of 3.7%, 5.7%, and 1.2% over the strongest
 317 CoT-RL baseline Open-Reasoner-Zero. Across model scales, UR² shows consistent advantages:
 318 on Qwen-2.5-3B, it achieves even larger performance gains with 9.1% improvement on MMLU-Pro
 319 and 8.6% on Medicine over Vanilla RL, demonstrating that UR² provides greater benefits for models
 320 with limited knowledge but strong reasoning capabilities. On LLaMA-3.1-8B, it achieves 43.4% on
 321 MMLU-Pro, outperforming all baselines. Notably, our method achieves competitive performance
 322 with the more capable closed-source GPT-4o-mini model on several tasks. As shown in Tables 1
 323 and 6, advanced RAG methods degrade performance on smaller models and require unacceptable
 324 source consumption (except Search-o1).
 325

³<https://huggingface.co/BAAI/bge-large-en-v1.5>

⁴<https://nlp.stanford.edu/projects/hotpotqa/enwiki-20171001-pages-meta-current-withlinks-abstracts.tar.bz2>

Table 1: Performance on reasoning and math tasks. We report EM scores (in %) on MMLU-Pro and MedQA, and LLM-as-a-judge scores (in %) on math benchmarks. \dagger = in-domain, \ddagger = out-of-domain. Best results are **bold**; second-best are underlined.

| Method | MMLU-Pro | | | | | Medicine | | | Math | | |
|------------------------------|-----------------|-----------------|-----------------|----------------|-------------|-----------------|------------------|-------------|-------------------|--------------------|-------------|
| | Hist. \dagger | Phil. \dagger | Econ. \dagger | Law \ddagger | Avg | MedQA \dagger | M-Med \ddagger | Avg | Math500 \dagger | Minerva \ddagger | Avg |
| GPT-4o-mini | | | | | | | | | | | |
| CoT | 56.7 | <u>53.1</u> | <u>70.4</u> | 38.2 | <u>54.5</u> | 71.4 | 67.0 | 69.2 | <u>78.0</u> | 65.6 | 71.8 |
| Standard RAG | <u>57.0</u> | 52.3 | 68.6 | <u>36.2</u> | 53.5 | 70.6 | 64.2 | 67.4 | 77.1 | 68.4 | <u>72.8</u> |
| <i>Advanced RAG Methods</i> | | | | | | | | | | | |
| Self-Ask | 56.3 | 48.5 | 67.8 | 31.2 | 51.0 | 72.4 | <u>68.0</u> | 70.2 | 62.9 | 45.2 | 54.1 |
| RAT | 57.5 | 55.3 | 73.0 | 34.2 | 55.0 | 74.4 | 70.6 | 72.5 | 77.5 | 64.2 | 70.9 |
| Search-o1 | 53.5 | 55.3 | 69.8 | 35.4 | 53.5 | 75.2 | 66.6 | <u>70.9</u> | 78.6 | <u>68.3</u> | 73.5 |
| Qwen-2.5-7B | | | | | | | | | | | |
| CoT | 42.3 | 45.7 | 63.4 | 26.6 | 44.5 | 57.2 | 52.0 | 54.6 | 76.6 | 59.4 | 68.0 |
| Standard RAG | 44.6 | 41.1 | 57.8 | 26.0 | 42.4 | 54.2 | 53.2 | 53.7 | 73.8 | 54.6 | 64.2 |
| <i>Advanced RAG Methods</i> | | | | | | | | | | | |
| Self-Ask | 40.7 | 42.1 | 60.0 | 26.2 | 42.3 | 51.8 | 47.8 | 49.8 | 74.9 | 57.7 | 66.3 |
| RAT | 47.2 | 44.7 | 64.4 | 30.0 | 46.6 | 60.0 | 53.2 | 56.6 | 74.4 | 55.5 | 65.0 |
| Search-o1 | 42.8 | 45.9 | 63.2 | 29.6 | 45.4 | 58.2 | 52.8 | 55.6 | 78.2 | 60.3 | 69.3 |
| <i>CoT-RL Methods</i> | | | | | | | | | | | |
| General Reasoner | 47.9 | 44.2 | 65.9 | 30.4 | 47.1 | 58.4 | 54.4 | 56.4 | 76.6 | 62.1 | <u>69.8</u> |
| Open-Reasoner-Zero | 50.0 | <u>46.6</u> | <u>67.5</u> | <u>34.2</u> | <u>49.6</u> | 61.6 | <u>58.8</u> | 60.2 | <u>80.7</u> | 58.8 | <u>69.8</u> |
| SimpleRL-Zoo | 35.7 | 36.9 | 55.2 | 25.4 | 38.3 | 57.2 | 51.0 | 54.1 | 77.1 | 50.7 | 63.9 |
| <i>Our Implementations</i> | | | | | | | | | | | |
| Vanilla RL | <u>52.2</u> | 43.5 | 64.0 | 33.8 | 48.4 | <u>64.2</u> | 57.4 | <u>60.8</u> | 78.2 | 59.4 | 68.8 |
| UR² (Ours) | 53.2 | 53.0 | 72.2 | 35.0 | 53.3 | 69.6 | 62.8 | 66.2 | 80.9 | <u>61.0</u> | 71.0 |
| Qwen-2.5-3B | | | | | | | | | | | |
| CoT | 33.6 | 32.3 | 48.8 | 20.6 | 33.8 | 39.4 | 36.8 | 38.1 | 63.6 | 39.9 | 51.8 |
| Standard RAG | 37.8 | 36.5 | 51.4 | 23.2 | 37.1 | 45.6 | 40.0 | 42.8 | 65.3 | 40.8 | 53.1 |
| <i>Our Implementations</i> | | | | | | | | | | | |
| Vanilla RL | 40.7 | 34.7 | 55.0 | 24.6 | 38.7 | 51.8 | 47.6 | 49.7 | 68.0 | 43.9 | 56.0 |
| UR² (Ours) | 47.8 | 49.3 | 63.9 | 30.0 | 47.8 | 59.8 | 56.8 | 58.3 | 69.4 | 45.0 | 57.2 |
| LLaMA-3.1-8B | | | | | | | | | | | |
| CoT | 37.8 | 40.9 | <u>53.4</u> | 29.0 | 40.3 | 59.6 | 52.6 | 56.1 | 48.4 | 34.4 | 41.4 |
| Standard RAG | 43.6 | 33.9 | 51.0 | 26.6 | 38.8 | 56.4 | 53.2 | 54.8 | 45.0 | 31.4 | 38.2 |
| <i>Our Implementations</i> | | | | | | | | | | | |
| Vanilla RL | <u>44.6</u> | 36.9 | 53.0 | 26.4 | 40.2 | <u>66.8</u> | <u>57.4</u> | 62.1 | <u>45.5</u> | 43.4 | <u>44.4</u> |
| UR² (Ours) | 48.3 | <u>38.6</u> | 58.0 | <u>28.8</u> | 43.4 | 68.6 | 58.4 | 63.5 | 54.5 | 39.0 | 46.8 |

5.2 MAIN RESULTS ON OPEN-DOMAIN QA

As demonstrated in Table 2, UR² achieves strong performance on open-domain QA, with Qwen-2.5-7B reaching 58.5% average F1 score, outperforming the strongest RAG-RL baseline Search-R1 (56.1%) by 2.4%. UR² demonstrates particularly robust out-of-domain generalization, achieving 64.5% on Bamboogle and 35.8% on Musique, surpassing all baselines. Across model scales, UR² maintains consistent advantages: on Qwen-2.5-3B, it achieves 55.3% F1, improving 8.2% over Searh-R1; on LLaMA-3.1-8B, it reaches 56.3%, competitive with specialized RAG-RL methods. Notably, our 7B model surpasses GPT-4.1-mini (55.4%) by 3.1%, demonstrating UR²'s effective dynamic coordination of retrieval and reasoning.

6 FURTHER ANALYSIS

Additional experimental results are provided in the Appendix, including further ablation studies (Appendix B.2), the impact of LLM summaries and corpus on UR² performance (Appendix B.3), comparative analysis of retrieval integration in RL training (Appendix B.4), unsuccessful attempts on reasoning models (Appendix B.5), and illustrative case studies (Appendix E).

378 Table 2: Performance on open-domain QA tasks. We report F1 and LLM-as-a-judge (LSJ)
 379 scores, both in %. \dagger = in-domain; \ddagger = out-of-domain.

| 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 Models | 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 Types | 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 Methods | 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 Hotpot \dagger | | 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 2Wiki \dagger | | 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 Bamb. \ddagger | | 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 MusiQ. \ddagger | | 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 Avg | |
|---|---|--|--|-------------|---|-------------|--|-------------|---|-------------|---|-------------|
| F1 | LSJ | F1 | LSJ | F1 | LSJ | F1 | LSJ | F1 | LSJ | F1 | LSJ | |
| 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 GPT-4.1-mini | Vanilla Methods | CoT | 43.7 | 59.2 | 48.6 | 60.8 | 59.2 | 76.0 | 28.3 | 35.4 | 45.0 | 57.9 |
| | | Standard RAG | 54.5 | 74.4 | 41.3 | 52.4 | 46.4 | 51.2 | 21.9 | 28.4 | 41.0 | 51.6 |
| | Advanced RAG | Self-Ask | 65.4 | 75.0 | 52.7 | 57.4 | 71.7 | 75.2 | 31.6 | 35.0 | 55.4 | 60.7 |
| | | RAT | 56.9 | 64.2 | 45.7 | 49.0 | 60.3 | 62.4 | 29.0 | 31.4 | 48.0 | 51.8 |
| | | Search-o1 | 53.1 | 74.0 | 44.4 | 60.6 | 63.7 | 71.2 | 28.6 | 33.4 | 47.5 | 59.8 |
| | | Vanilla Methods | CoT | 24.9 | 31.0 | 25.1 | 27.6 | 41.3 | 43.2 | 14.8 | 12.2 | 26.5 |
| | Qwen-2.5-7B | Standard RAG | 49.2 | 62.8 | 32.8 | 37.6 | 38.9 | 40.0 | 14.4 | 14.6 | 33.8 | 38.8 |
| | | Self-Ask | 28.8 | 61.0 | 22.2 | 45.4 | 28.9 | 42.4 | 13.6 | 19.6 | 23.4 | 42.1 |
| | | Advanced RAG | 37.9 | 40.6 | 23.3 | 23.6 | 31.6 | 30.4 | 14.4 | 12.4 | 26.8 | 26.8 |
| | | Search-o1 | 50.9 | 61.6 | 45.2 | 48.6 | 37.5 | 39.2 | 20.6 | 19.8 | 38.6 | 42.3 |
| | | R1-Searcher | 71.8 | 78.0 | 57.9 | 63.6 | 56.5 | 53.6 | 33.3 | 32.6 | 54.8 | 57.0 |
| | | Search-R1 | 72.4 | 78.8 | 61.0 | 63.8 | 58.9 | 56.8 | 32.2 | 32.0 | 56.1 | 57.9 |
| | RAG-RL | R1-Searcher++ | 59.0 | 64.2 | 61.2 | 64.4 | 60.8 | 59.2 | 33.8 | 32.8 | 53.7 | 55.2 |
| | | ZeroSearch | 46.0 | 50.4 | 38.4 | 38.6 | 35.8 | 38.4 | 14.7 | 13.8 | 33.7 | 35.3 |
| | | Vanilla RL | 70.9 | 78.8 | 61.2 | 62.4 | 63.3 | 63.2 | 34.4 | 34.4 | 57.5 | 59.6 |
| | | UR² (Ours) | 71.2 | 79.4 | 62.6 | 65.0 | 64.5 | 62.4 | 35.8 | 34.6 | 58.5 | 60.4 |
| 401 402 403 404 405 406 407 408 409 410 411 412 Qwen-2.5-3B | Vanilla Methods | CoT | 26.6 | 27.2 | 22.7 | 22.6 | 31.2 | 33.6 | 11.3 | 9.6 | 23.0 | 23.3 |
| | | Standard RAG | 50.6 | 57.0 | 29.8 | 30.4 | 26.1 | 27.2 | 9.7 | 7.4 | 29.1 | 30.5 |
| | RAG-RL | Search-R1 | 63.1 | 69.2 | 49.5 | 53.4 | 48.3 | 48.0 | 27.6 | 27.8 | 47.1 | 49.6 |
| | | Zero-Search | 42.7 | 45.8 | 26.1 | 27.6 | 32.4 | 31.2 | 16.9 | 17.0 | 29.5 | 30.4 |
| | | Vanilla RL | 65.9 | 73.6 | 54.9 | 58.0 | 59 | 57.6 | 30.0 | 29.6 | 52.5 | 54.7 |
| | | UR² (Ours) | 67.7 | 76.0 | 55.2 | 58.6 | 57.8 | 58.4 | 30.5 | 31.6 | 55.3 | 56.2 |
| | LLaMA-3.1-8B | Vanilla Methods | 28.6 | 31.6 | 16.4 | 17.8 | 43.0 | 42.4 | 9.8 | 10.8 | 24.5 | 25.7 |
| | | Standard RAG | 47.5 | 54.4 | 26.2 | 26.4 | 26.5 | 28.0 | 10.1 | 10.2 | 27.6 | 29.8 |
| | | R1-Searcher | 70.8 | 76.8 | 59.6 | 62.2 | 64.7 | 62.4 | 31.1 | 29.4 | 56.6 | 57.7 |
| | | Vanilla RL | 70.0 | 77.6 | 61.2 | 64.2 | 60.6 | 63.2 | 32.7 | 31.8 | 56.1 | 59.2 |
| | Our Implementations | UR² (Ours) | <u>70.1</u> | 78.8 | <u>60.1</u> | <u>63.2</u> | <u>60.7</u> | 63.2 | 34.3 | 34.0 | <u>56.3</u> | 59.8 |

6.1 IMPACT OF ONLINE SEARCH

To test scalability under online retrieval, we compare local corpus with real-time search (Table 3). Online search yields consistent gains on MMLU-Pro and medical tasks, and substantial improvements on 2Wiki and Bamboogle, demonstrating strong generalization to scenarios requiring up-to-date or non-Wikipedia knowledge. The only exceptions are math—where Qwen-3-32B’s parametric knowledge already covers the required formulas and axioms—and HotpotQA, where rate

Table 3: Comparison of UR² Qwen-2.5-7B-Instruct using Local Corpus vs. Online Search across different tasks. \dagger = in-domain; \ddagger = out-of-domain.

| 423 424 425 426 427 428 429 430 431 Corpus | 423 424 425 426 427 428 429 430 431 MMLU-Pro | | | | | 423 424 425 426 427 428 429 430 431 Medicine | | | | | 423 424 425 426 427 428 429 430 431 Math | | | | | | | | |
|---|---|-----------------|-----------------|----------------|-------------|---|------------------|-------------|-------------------|--------------------|---|------|-----|-----|-----|-------------------|-----|-----|-----|
| | Hist. \dagger | Phil. \dagger | Econ. \dagger | Law \ddagger | Avg | MedQA \dagger | M-Med \ddagger | Avg | Math500 \dagger | Minerva \ddagger | Avg | F1 | LSJ | F1 | LSJ | F1 | LSJ | | |
| Local Corpus | 53.2 | 53.0 | 72.2 | 35.0 | 53.3 | 69.6 | 62.8 | 65.9 | 80.9 | 61.0 | 71.0 | | | | | | | | |
| Online Search | 57.7 | 57.8 | 71.0 | 35.0 | 55.4 | 70.4 | 65.4 | 67.9 | 78.7 | | 61.2 | 70.0 | | | | | | | |
| Corpus | Hotpot \dagger | | | | | 2Wiki \dagger | | | | | Bamb. \ddagger | | | | | MusiQ. \ddagger | | Avg | |
| | F1 | LSJ | F1 | LSJ | F1 | F1 | LSJ | F1 | LSJ | F1 | F1 | LSJ | F1 | LSJ | F1 | F1 | LSJ | F1 | LSJ |
| Local Corpus | 71.2 | 79.4 | 62.6 | 65.0 | 64.5 | 62.4 | 35.8 | 34.6 | 58.5 | 60.4 | | | | | | | | | |
| Online Search | 62.0 | 67.6 | 75.8 | 81.8 | 73.7 | 76.0 | 34.9 | 37.8 | 61.6 | 65.8 | | | | | | | | | |

limits block access to many gold Wikipedia pages. Notably, our setting does not enforce full top-10 coverage, which naturally introduces noise and better reflects real-world retrieval conditions. Overall, these results confirm the robustness of our method in noisy online environments.

6.2 ABLATION STUDY

To validate the effectiveness of UR², we conduct comprehensive ablation studies on its key components. As shown in Table 4, all variants exhibit performance degradation. The *W/o Stage-1* variant causes notable drops (5.2% in History, 4.2% in Economics), demonstrating that explicit retrieval activation is essential. The *W/o P_{fallback}* variant shows slight improvements on Law and MedQA but generates unreasonable queries, such as “**which option is right**”. The *W/o LLM Summary* variant completely fails, as models degrade to pure CoT, **highlighting the necessity of addressing retrieval noise in RAG-RL methods**. The *W/o Task Mixing* variant shows minimal changes, indicating that our selective strategy **improves efficiency while enhancing accuracy**. Using alternative LLMs for corpus summarization (4omini/Qw3-8B) results in consistent 3-4% drops across tasks but still outperforms vanilla RL, demonstrating our method’s adaptability in resource-constrained settings. These results show that our two-stage training, difficulty-aware retrieval, and carefully designed reward components work synergistically to achieve superior performance.

Table 4: Ablation study of Qwen-2.5-7B-Instruct on MMLU-Pro and medical reasoning tasks. “w/o Task Mixing” means retrieving for all samples. \dagger = in-domain; \ddagger = out-of-domain.

| Method | MMLU-Pro | | | | | Medicine | | |
|---------------------------|-----------------|-----------------|-----------------|----------------|-------------|-----------------|------------------|-------------|
| | Hist. \dagger | Phil. \dagger | Econ. \dagger | Law \ddagger | Avg | MedQA \dagger | M-Med \ddagger | Avg |
| UR ² | 53.2 | 53.1 | 72.2 | 35.0 | 53.3 | 69.6 | 62.8 | 66.2 |
| w/o Stage-1 | 48.0 | 51.1 | 68.0 | 30.9 | 49.5 | 67.6 | 63.0 | 65.3 |
| w/o P _{fallback} | 52.0 | 51.3 | 68.4 | 36.6 | 52.1 | 71.4 | 62.0 | 66.7 |
| w/o Task Mixing | 52.2 | 51.9 | 68.2 | 33.2 | 51.4 | 70.0 | 63.6 | 66.8 |
| w/o LLM Summary | — | — | — | — | — | — | — | — |
| Vanilla RL | 52.2 | 43.5 | 64.0 | 33.8 | 48.4 | 64.2 | 57.4 | 60.8 |
| 4omini Summary | 49.3 | 48.8 | 67.4 | 32.4 | 49.5 | 65.0 | 59.2 | 62.1 |
| Qw3-8B Summary | 49.1 | 49.9 | 67.8 | 30.6 | 49.4 | 64.8 | 58.2 | 61.5 |

Table 5 validates our difficulty-aware data selection strategy. Despite using significantly less data, filtered datasets achieve comparable or superior performance, particularly on out-of-domain tasks (Bamboolge improves from 58.9% to 62.7%). This confirms that RL training benefits more from high-quality, difficulty-balanced samples than large-scale unfiltered data, enabling computationally efficient training while maintaining strong performance across diverse tasks.

Table 5: Training Data Ablation for Qwen-2.5-7B-Instruct (Vanilla RL on Open-domain QA w/o Summary and Math Tasks). We report F1 scores (%) for open-domain QA.

| Method | HotpotQA \dagger | 2Wiki \dagger | Bamboolge \ddagger | MusiQue \ddagger | Math \dagger | Minerva \ddagger |
|---------------|--------------------|-----------------|----------------------|--------------------|----------------|--------------------|
| Raw Data | 72.3 | 61.9 | 58.9 | 36.1 | 79.0 | 60.5 |
| Filtered Data | 71.0 | 62.0 | 62.7 | 33.8 | 78.2 | 59.4 |

7 CONCLUSION

In this work, we presented UR², a unified framework that integrates retrieval-augmented generation with reasoning through reinforcement learning. Unlike existing RAG-RL approaches limited to specific domains, UR² demonstrates versatility across mathematical reasoning, medical QA, and open-domain tasks. Our key innovations—difficulty-aware curriculum learning and an LLM-summarized retrieval corpus—enable dynamic retrieval-reasoning coordination by learning *when* and *what* to retrieve based on problem difficulty, while preserving native reasoning capabilities. UR² represents a significant step toward adaptive AI systems that flexibly combine parametric knowledge with dynamic information access.

486 REFERENCES
487

- 488 Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Milli-
489 can, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al.
490 Improving language models by retrieving from trillions of tokens. In *International conference on
491 machine learning*, pp. 2206–2240. PMLR, 2022.
- 492 Zhipeng Chen, Yingqian Min, Beichen Zhang, Jie Chen, Jinhao Jiang, Daixuan Cheng, Wayne Xin
493 Zhao, Zheng Liu, Xu Miao, Yang Lu, et al. An empirical study on eliciting and improving r1-like
494 reasoning models. *arXiv preprint arXiv:2503.04548*, 2025.
- 495 Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser,
496 Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to
497 solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- 498 Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha
499 Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models.
500 *arXiv e-prints*, pp. arXiv–2407, 2024.
- 501 Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt,
502 and Jonathan Larson. From local to global: A graph rag approach to query-focused summariza-
503 tion. *arXiv preprint arXiv:2404.16130*, 2024.
- 504 Tianyu Gao, Zhiyuan Liu, Xu Han, Ningyu Zhang, Jilin Tang, Maosong Sun, et al. Retrieval-
505 augmented generation for large language models: A survey. *arXiv preprint arXiv:2302.05100*,
506 2023.
- 507 Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu,
508 Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms
509 via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025a.
- 510 Yiduo Guo, Zhen Guo, Chuanwei Huang, Zi-Ang Wang, Zekai Zhang, Haofei Yu, Huishuai
511 Zhang, and Yikang Shen. Synthetic data rl: Task definition is all you need. *arXiv preprint
512 arXiv:2505.17063*, 2025b.
- 513 Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song,
514 and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv
515 preprint arXiv:2103.03874*, 2021.
- 516 Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-
517 hop QA dataset for comprehensive evaluation of reasoning steps. In Donia Scott, Nuria Bel,
518 and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational
519 Linguistics*, pp. 6609–6625, Barcelona, Spain (Online), December 2020. International Com-
520 mittee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.580. URL <https://aclanthology.org/2020.coling-main.580/>.
- 521 Jian Hu, Xibin Wu, Zilin Zhu, Xianyu, Weixun Wang, Dehao Zhang, and Yu Cao. Openrlhf: An
522 easy-to-use, scalable and high-performance rlhf framework. *arXiv preprint arXiv:2405.11143*,
523 2024.
- 524 Jingcheng Hu, Yinmin Zhang, Qi Han, Dixin Jiang, Xiangyu Zhang, and Heung-Yeung Shum.
525 Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base
526 model. *arXiv preprint arXiv:2503.24290*, 2025a.
- 527 Yuntong Hu, Zhihan Lei, Zheng Zhang, Bo Pan, Chen Ling, and Liang Zhao. GRAG: Graph
528 retrieval-augmented generation. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings
529 of the Association for Computational Linguistics: NAACL 2025*, pp. 4145–4157, Albuquerque,
530 New Mexico, April 2025b. Association for Computational Linguistics. ISBN 979-8-89176-195-
531 7. URL <https://aclanthology.org/2025.findings-naacl.232/>.
- 532 Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane
533 Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Few-shot learning with re-
534 trieval augmented language models. *arXiv preprint arXiv:2208.03299*, 2022.

- 540 Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and
 541 Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement
 542 learning. *arXiv preprint arXiv:2503.09516*, 2025.
- 543
- 544 Di Jin, Eileen Pan, Nassim Oufattolle, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What dis-
 545 ease does this patient have? a large-scale open domain question answering dataset from medical
 546 exams. *Applied Sciences*, 11(14):6421, 2021.
- 547
- 548 Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large
 549 language models are zero-shot reasoners. *Advances in neural information processing systems*,
 35:22199–22213, 2022.
- 550
- 551 Ezgi Korkmaz. Understanding and diagnosing deep reinforcement learning. *arXiv preprint*
 552 *arXiv:2406.16979*, 2024.
- 553
- 554 Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal,
 Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented genera-
 555 tion for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:
 556 9459–9474, 2020.
- 557
- 558 Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ra-
 559 masesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative
 560 reasoning problems with language models. *Advances in neural information processing systems*,
 35:3843–3857, 2022.
- 561
- 562 Weitao Li, Junkai Li, Weizhi Ma, and Yang Liu. Citation-enhanced generation for LLM-based
 563 chatbots. In Lun-Wei Ku, Andre Martins, and Vivek Srikanth (eds.), *Proceedings of the 62nd*
564 Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.
 565 1451–1466, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi:
 566 10.18653/v1/2024.acl-long.79. URL <https://aclanthology.org/2024.acl-long.79/>.
- 567
- 568 Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and
 569 Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models. *arXiv preprint*
 570 *arXiv:2501.05366*, 2025.
- 571
- 572 Xueguang Ma, Qian Liu, Dongfu Jiang, Ge Zhang, Zejun Ma, and Wenhui Chen. General-reasoner:
 573 Advancing llm reasoning across all domains. *arXiv preprint arXiv:2505.14652*, 2025.
- 574
- 575 Prabhat Nagarajan, Garrett Warnell, and Peter Stone. Deterministic implementations for repro-
 576 ducibility in deep reinforcement learning. *arXiv preprint arXiv:1809.05676*, 2018.
- 577
- 578 Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and
 579 Siliang Tang. Graph retrieval-augmented generation: A survey. *arXiv preprint arXiv:2408.08921*,
 2024.
- 580
- 581 Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James
 582 Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. Kilt: a benchmark for knowl-
 583 edge intensive language tasks. In *Proceedings of the 2021 Conference of the North American*
584 Chapter of the Association for Computational Linguistics: Human Language Technologies, pp.
 2523–2544, 2021.
- 585
- 586 Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. Measuring
 587 and narrowing the compositionality gap in language models. In *Findings of the Association for*
588 Computational Linguistics: EMNLP 2023, pp. 5687–5711, 2023.
- 589
- 590 Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations
 591 toward training trillion parameter models. In *SC20: International Conference for High Perfor-*
592 mance Computing, Networking, Storage and Analysis, pp. 1–16. IEEE, 2020.
- 593
- Jingyi Ren, Yekun Xu, Xiaolong Wang, Weitao Li, Weizhi Ma, and Yang Liu. Effective and trans-
 parent rag: Adaptive-reward reinforcement learning for decision traceability. *arXiv preprint*
arXiv:2505.13258, 2025.

- 594 John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy
 595 optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- 596
- 597 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang,
 598 Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical
 599 reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- 600 Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang,
 601 and Ji-Rong Wen. R1-searcher: Incentivizing the search capability in llms via reinforcement
 602 learning. *arXiv preprint arXiv:2503.05592*, 2025a.
- 603
- 604 Huatong Song, Jinhao Jiang, Wenqing Tian, Zhipeng Chen, Yuhuan Wu, Jiahao Zhao, Yingqian Min,
 605 Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. R1-searcher++: Incentivizing the dynamic knowl-
 606 edge acquisition of llms via reinforcement learning. *arXiv preprint arXiv:2505.17005*, 2025b.
- 607 Hao Sun, Hengyi Cai, Bo Wang, Yingyan Hou, Xiaochi Wei, Shuaiqiang Wang, Yan Zhang, and
 608 Dawei Yin. Towards verifiable text generation with evolving memory and self-reflection. In
 609 Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference
 610 on Empirical Methods in Natural Language Processing*, pp. 8211–8227, Miami, Florida, USA,
 611 November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.
 612 469. URL <https://aclanthology.org/2024.emnlp-main.469/>.
- 613 Hao Sun, Zile Qiao, Jiayan Guo, Xuanbo Fan, Yingyan Hou, Yong Jiang, Pengjun Xie, Fei Huang,
 614 and Yan Zhang. Zerosearch: Incentivize the search capability of llms without searching. *arXiv
 615 preprint arXiv:2505.04588*, 2025.
- 616
- 617 Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Mul-
 618 tihop questions via single-hop question composition. *Transactions of the Association for
 619 Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL <https://aclanthology.org/2022.tacl-1.31/>.
- 620
- 621 Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming
 622 Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-
 623 task language understanding benchmark. *Advances in Neural Information Processing Systems*,
 624 37:95266–95290, 2024a.
- 625
- 626 Zihao Wang, Anji Liu, Haowei Lin, Jiaqi Li, Xiaojian Ma, and Yitao Liang. Rat: Retrieval
 627 augmented thoughts elicit context-aware reasoning in long-horizon generation. *arXiv preprint
 628 arXiv:2403.05313*, 2024b.
- 629
- 630 Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement
 631 learning. *Machine learning*, 8:229–256, 1992.
- 632
- 633 Guangzhi Xiong, Qiao Jin, Xiao Wang, Yin Fang, Haolin Liu, Yifan Yang, Fangyuan Chen, Zhixing
 634 Song, Dengyu Wang, Minjia Zhang, et al. Rag-gym: Optimizing reasoning and search agents
 635 with process supervision. *arXiv preprint arXiv:2502.13957*, 2025.
- 636
- 637 An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li,
 638 Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv e-prints*, pp.
 639 arXiv–2412, 2024.
- 640
- 641 Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov,
 642 and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question
 643 answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language
 644 Processing*, pp. 2369–2380, 2018.
- 645
- 646 Qiyi Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian
 647 Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system
 648 at scale. *arXiv preprint arXiv:2503.14476*, 2025.
- 649
- 650 Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. Simplerl-
 651 zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv
 652 preprint arXiv:2503.18892*, 2025.

648 Xiaoying Zhang, Hao Sun, Yipeng Zhang, Kaituo Feng, Chao Yang, and Helen Meng. Critique-
 649 grp: Advancing llm reasoning with natural language and numerical feedback. *arXiv preprint*
 650 *arXiv:2506.03106*, 2025.

651 Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei
 652 Liu. Deepresearcher: Scaling deep research via reinforcement learning in real-world environ-
 653 ments. *arXiv preprint arXiv:2504.03160*, 2025.

656 A LIMITATIONS, FUTURE DIRECTIONS, AND AI USAGE

658 **Limitations.** While UR² shows strong performance across diverse tasks, some limitations remain.
 659 First, we have not scaled beyond 8B parameters due to computational limits. Second, our reliance on
 660 LLM-summarized corpora may not fully reflect the complexity of raw web content. Nevertheless,
 661 the use of offline or noise-reduced corpora is a common practice in prior works (Zheng et al., 2025;
 662 Song et al., 2025a; Ren et al., 2025; Sun et al., 2025; Jin et al., 2025).

663 **Future Directions.** Future work will explore updated models and frameworks, scaling UR² to 32B
 664 parameters, and incorporating online corpora during training to better capture real-world retrieval
 665 dynamics. We also plan to investigate more efficient training strategies to reduce costs.

667 **Use of AI Tools.** In preparing this work, we used commercial LLMs (e.g., Claude 4.0) for non-
 668 creative assistance such as language polishing, formatting, and minor code edits. These tools were
 669 not involved in method design, experimental setup, or any substantive creative contribution.

671 B ADDITIONAL EXPERIMENTAL RESULTS

673 B.1 COMPREHENSIVE SUPPLEMENTARY RESULTS ON OPEN-DOMAIN AND REASONING 674 TASKS

676 Tables 6 and 7 provide supplementary experimental results, focusing on Advanced RAG methods
 677 across different model scales and GPT-4o-mini performance on open-domain QA tasks.

678 Table 6: Extended results on GPT-4o-mini, Qwen-2.5-3B-Instruct, and LLaMA-3.1-8B-Instruct
 679 across reasoning tasks. We report EM scores (%) for MMLU-Pro and MedQA, and
 680 LLM-as-a-judge scores (%) for math benchmarks. † = in-domain; ‡ = out-of-domain.

| Method | MMLU-Pro | | | | | Medicine | | | | Math | | |
|-----------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|--|
| | Hist.† | Phil.† | Econ.† | Law‡ | Avg | MedQA† | M-Med‡ | Avg | Math500† | Minerva‡ | Avg | |
| <i>Qwen-2.5-3B</i> | | | | | | | | | | | | |
| CoT | 33.6 | 32.3 | 48.8 | 20.6 | 33.8 | 39.4 | 36.8 | 38.1 | 63.6 | 39.9 | 51.8 | |
| Standard RAG | 37.8 | <u>36.5</u> | 51.4 | 23.2 | 37.1 | 45.6 | 40.0 | 42.8 | 65.3 | 40.8 | 53.1 | |
| <i>Advanced RAG Methods</i> | | | | | | | | | | | | |
| Self-Ask | 33.1 | 30.5 | 44.2 | 20.2 | 32.0 | 39.6 | 36.0 | 37.8 | 58.3 | 40.3 | 49.3 | |
| RAT | 37.8 | 33.5 | 50.6 | 18.4 | 35.1 | 47.4 | 42.6 | 45.0 | 67.9 | <u>44.9</u> | <u>56.4</u> | |
| Search-o1 | 33.9 | 34.5 | 50.6 | 20.6 | 34.9 | 44.6 | 37.4 | 41.0 | 69.4 | 43.0 | 56.2 | |
| <i>Our Implementations</i> | | | | | | | | | | | | |
| Vanilla RL | <u>40.7</u> | 34.7 | <u>55.0</u> | <u>24.6</u> | <u>38.7</u> | <u>51.8</u> | <u>47.6</u> | <u>49.7</u> | 68.0 | 43.9 | 52.4 | |
| UR² | 47.8 | 49.3 | 63.9 | 30.0 | 47.8 | 59.8 | 56.8 | 58.3 | 69.4 | 45.0 | 57.2 | |
| <i>LLaMA-3.1-8B</i> | | | | | | | | | | | | |
| CoT | 37.8 | 40.9 | <u>53.4</u> | 29.0 | 40.3 | 59.6 | 52.6 | 56.1 | 48.4 | 34.4 | 41.4 | |
| Standard RAG | <u>43.6</u> | 33.9 | 51.0 | 26.6 | 38.8 | 56.4 | 53.2 | 54.8 | 45.0 | 31.4 | 38.2 | |
| <i>Advanced RAG Methods</i> | | | | | | | | | | | | |
| Self-Ask | 39.8 | 32.1 | 47.0 | 23.4 | 35.6 | 53.0 | 42.8 | 47.9 | 46.9 | 27.0 | 37.0 | |
| RAT | 42.3 | 37.7 | 52.6 | 28.6 | 40.3 | 63.8 | 56.0 | 59.9 | <u>50.1</u> | 36.8 | 43.5 | |
| Search-o1 | 32.6 | 32.5 | 46.0 | 28.0 | 45.9 | 56.0 | 46.0 | 56.6 | 41.5 | 27.8 | 34.7 | |
| <i>Our Implementations</i> | | | | | | | | | | | | |
| Vanilla RL | 44.6 | 36.9 | 53.0 | 26.4 | 40.2 | <u>66.8</u> | <u>57.4</u> | <u>62.1</u> | 45.5 | 43.4 | <u>44.4</u> | |
| UR² | 48.3 | <u>38.6</u> | 58.0 | <u>28.8</u> | <u>43.4</u> | 68.6 | 58.4 | <u>63.5</u> | <u>54.5</u> | <u>39.0</u> | 46.8 | |

The extended results reveal significant performance limitations of Advanced RAG methods for open-source models. On Qwen-2.5-3B, Self-Ask achieves only 32.0% on MMLU-Pro, substantially underperforming even basic CoT (33.8%). RAT shows inconsistent performance, achieving competitive results on medical tasks (45.0%) but poor performance on Law (18.4%), indicating fragility in cross-domain generalization. Search-o1 demonstrates moderate effectiveness, reaching 41.0% on medical tasks, but fails to achieve consistent improvements across reasoning domains. On LLaMA-3.1-8B, Advanced RAG methods exhibit mixed results. While RAT achieves reasonable performance on Medicine (59.9%) and Math (43.5%), Self-Ask and Search-o1 show notable degradation compared to basic CoT on several sub-domains. These results highlight the challenge of scaling sophisticated retrieval mechanisms to diverse model architectures and reasoning tasks.

GPT-4o-mini establishes strong performance on open-domain QA, with Search-o1 achieving 48.9% F1 average, significantly outperforming other Advanced RAG methods (41.3 and 41.8%). Additionally, RAT and Self-Ask incur prohibitive API costs due to their sentence-level analysis and rewriting operations, making them impractical for large-scale deployment. Notably, Standard RAG achieves competitive performance (42.1% F1) on GPT-4o-mini, suggesting that larger commercial models can effectively leverage simple retrieval without sophisticated coordination mechanisms. The performance gap between GPT-4o-mini (48.9%) and smaller models, such as Qwen-2.5-3B (27.8%) for Search-o1, highlights the substantial challenge of achieving effective retrieval-reasoning integration in resource-constrained settings and validates the necessity of our specialized framework design.

Table 7: Extended results of GPT-4o-mini, Qwen-2.5-3B, and LLaMA-3.1-8B on open-domain QA. We report F1 and LLM-as-a-judge (LSJ) scores, both in %. \dagger denotes in-domain datasets; \ddagger indicates out-of-domain.

| Models | Types | Methods | Hotpot † | | 2Wiki † | | Bamb. ‡ | | MusiQ. ‡ | | Avg | |
|---------------------|---------------------|-----------------------|-------------------|-------------|------------------|-------------|-------------------|-------------|--------------------|-------------|-------------|-------------|
| | | | F1 | LSJ | F1 | LSJ | F1 | LSJ | F1 | LSJ | | |
| GPT-4o-mini | Vanilla Methods | CoT | 46.5 | 51.2 | 35.0 | 35.4 | 55.2 | 62.4 | 24.9 | 26.8 | 40.4 | 44.0 |
| | | Standard RAG | 59.6 | 69.6 | 43.0 | 45.8 | 46.7 | 46.4 | 19.3 | 21.6 | 42.1 | 45.9 |
| | Advanced RAG | Self-Ask | 45.0 | 50.4 | 36.9 | 40.0 | 59.3 | 57.6 | 26.1 | 27.8 | 41.8 | 44.0 |
| | | RAT | 53.8 | 59.2 | 34.1 | 34.8 | 53.0 | 51.2 | 24.3 | 24.8 | 41.3 | 42.5 |
| | | Search-o1 | 64.3 | 73.4 | 47.3 | 52.0 | 54.5 | 56.0 | 29.6 | 30.2 | 48.9 | 52.9 |
| | | CoT | 26.6 | 27.2 | 22.7 | 22.6 | 31.2 | 33.6 | 11.3 | 9.6 | 23.0 | 23.3 |
| Qwen-2.5-3B | Vanilla Methods | Standard RAG | 50.6 | 57.0 | 29.8 | 30.4 | 26.1 | 27.2 | 9.7 | 7.4 | 29.1 | 30.5 |
| | | Self-Ask | 33.8 | 47.2 | 21.0 | 28.8 | 30.6 | 32.0 | 14.5 | 14.8 | 25.0 | 30.7 |
| | Advanced RAG | RAT | 30.1 | 32.2 | 15.1 | 15.4 | 30.6 | 28.0 | 11.0 | 8.2 | 21.7 | 21.0 |
| | | Search-o1 | 36.4 | 37.6 | 30.8 | 31.8 | 31.4 | 32.0 | 12.5 | 10.0 | 27.8 | 27.9 |
| | | Search-R1 | 63.1 | 69.2 | 49.5 | 53.4 | 48.3 | 48.0 | 27.6 | 27.8 | 49.6 | 47.1 |
| | | Zero-Search | 42.7 | 45.8 | 26.1 | 27.6 | 32.4 | 31.2 | 16.9 | 17.0 | 29.5 | 30.4 |
| | Our Implementations | Vanilla RL | 65.9 | 73.6 | 54.9 | 58.0 | 59 | 57.6 | 30.0 | 29.6 | 52.5 | 54.7 |
| | | UR² | 67.7 | 76.0 | 55.2 | 58.6 | 57.8 | 58.4 | 30.5 | 31.6 | 55.3 | 56.2 |
| LLaMA-3.1-8B | Vanilla Methods | CoT | 28.6 | 31.6 | 16.4 | 17.8 | 43.0 | 42.4 | 9.8 | 10.8 | 24.5 | 25.7 |
| | | Standard RAG | 47.5 | 54.4 | 26.2 | 26.4 | 26.5 | 28.0 | 10.1 | 10.2 | 27.6 | 29.8 |
| | Advanced RAG | Self-Ask | 43.0 | 50.8 | 27.3 | 29.8 | 41.5 | 44.8 | 16.8 | 16.4 | 32.2 | 35.5 |
| | | RAT | 44.5 | 48.8 | 16.4 | 15.6 | 39.7 | 39.2 | 17.0 | 16.0 | 29.4 | 29.9 |
| | | Search-o1 | 53.0 | 59.4 | 37.5 | 38.4 | 30.0 | 30.4 | 15.9 | 16.2 | 34.1 | 36.1 |
| | | RAG-RL | 70.8 | 76.8 | 59.6 | 62.2 | 64.7 | 62.4 | 31.1 | 29.4 | 56.6 | 57.7 |
| | Our Implementations | Vanilla RL | 70.0 | 77.6 | 61.2 | 64.2 | 60.6 | 63.2 | 32.7 | 31.8 | 56.1 | 59.2 |
| | | UR² | <u>70.1</u> | 78.8 | <u>60.1</u> | <u>63.2</u> | <u>60.7</u> | 63.2 | 34.3 | 34.0 | <u>56.3</u> | 59.8 |

B.2 ADDITIONAL ABLATION RESULTS

To ensure a fair comparison, we evaluate Vanilla RL MCQ (MMLU-Pro and Medicine tasks), which trains on mixed multiple-choice tasks similar to UR². As shown in Table 8, Vanilla RL MCQ exhibits task-dependent performance: on Qwen-2.5-7B it improves Medicine performance (62.8%

vs. 60.8%) but lowers MMLU-Pro scores (47.0% vs. 48.4%), with the reverse trend on 3B models. Despite these gains, UR² consistently outperforms both Vanilla RL variants across all domains and scales, achieving average improvements on MMLU-Pro of 5.9% for 7B models and 9.1% for 3B models, confirming that its advantage arises from the unified retrieval-reasoning framework rather than task mixing alone.

Table 8: Ablation study of Vanilla RL on Qwen-2.5-7B-Instruct and Qwen-2.5-3B-Instruct across multiple-choice reasoning tasks.

| Method | MMLU-Pro | | | | | Medicine | | |
|---------------------------|--------------------|--------------------|--------------------|------------------|-------------|--------------------|--------------------|-------------|
| | Hist. [†] | Phil. [†] | Econ. [†] | Law [‡] | Avg | MedQA [†] | M-Med [‡] | Avg |
| <i>Qwen-2.5-7B</i> | | | | | | | | |
| Vanilla RL MCQ | 47.2 | <u>46.1</u> | 61.8 | 33.0 | 47.0 | <u>65.6</u> | <u>60.0</u> | <u>62.8</u> |
| Vanilla RL | <u>52.2</u> | 43.5 | <u>64.0</u> | <u>33.8</u> | <u>48.4</u> | 64.2 | 57.4 | 60.8 |
| UR² | 53.2 | 53.0 | 72.2 | 35.0 | 53.3 | 69.6 | 62.8 | 65.9 |
| <i>Qwen-2.5-3B</i> | | | | | | | | |
| Vanilla RL MCQ | 42.3 | 37.1 | 57.4 | 25.0 | 40.6 | 50.2 | 45.0 | 47.6 |
| Vanilla RL | 40.7 | 34.7 | 55.0 | 24.6 | 38.7 | 51.8 | 47.6 | 49.7 |
| UR² | 47.8 | 49.3 | 63.9 | 30.0 | 47.8 | 59.8 | 56.8 | 58.3 |

We further conduct ablation studies on UR² in open-domain QA tasks (Table 9). The *w/o Math Data* variant shows minimal impact (0.3-1.4% drops), confirming multi-task training preserves QA performance. Additionally removing LLM Summary causes larger drops on out-of-domain tasks (2.0% on MusiQue) while maintaining in-domain performance, indicating LLM-summarized corpus benefits generalization. The *weaker Stage-1* variant shows the largest degradation on Bamboogle (5.5% drop), highlighting proper retrieval initialization is crucial for complex multi-hop reasoning. These results validate our design choices contribute meaningfully across diverse task types.

Table 9: Ablation Study of Qwen-2.5-7B-Instruct on open-domain QA. We report F1 scores (in %) here. The second variant removes LLM Summary on top of the first variant without Math Data.

| Method | Hotpot [†] | 2Wiki [†] | Bamb. [‡] | MusiQ. [‡] |
|-----------------------|---------------------|--------------------|--------------------|---------------------|
| UR² | 71.2 | 62.6 | 64.5 | 35.8 |
| w/o Math Data | 70.9 | 61.2 | 63.3 | 34.4 |
| w/o LLM Summary | 71.0 | 62.0 | 62.7 | 33.8 |
| weaker Stage-1 | 69.5 | 62.2 | 59.0 | 34.4 |

Table 10: Ablation on summarizers in UR² (Qwen-2.5-7B-Instruct) on MMLU-Pro. “w/o Summary” uses top-3 documents without summarizing; “Qwen-3-32B” uses top-16 documents; “Qwen-2.5-7B” (instruct) uses top-5 documents.

| Summarizer | MMLU-Pro (EM %) | | | | AVG |
|--------------|--------------------|--------------------|--------------------|------------------|-------------|
| | Hist. [†] | Phil. [†] | Econ. [†] | Law [‡] | |
| GPT-4.1 | <u>53.2</u> | 53.1 | 72.2 | 35.0 | 53.4 |
| Qwen-3-32B | 52.5 | 50.9 | <u>72.0</u> | 33.6 | <u>52.3</u> |
| Qwen-3-8B | 52.5 | <u>51.5</u> | 71.0 | 32.8 | 52.0 |
| GPT-4.1-mini | 52.5 | 51.3 | 69.4 | 33.6 | 51.7 |
| GPT-4o-mini | 51.8 | 49.1 | 67.4 | <u>34.5</u> | 50.8 |
| Qwen-2.5-7B | 53.4 | 47.4 | 67.2 | 32.0 | 50.0 |
| w/o Summary | 52.1 | 48.3 | 68.0 | 32.2 | 50.2 |
| Vanilla RL | 52.2 | 43.5 | 64.0 | 33.8 | 48.4 |

Overall, the ablations confirm that Stage-1 initialization is crucial for complex reasoning, difficulty-aware filtering yields better performance with fewer samples, and task mixing improves efficiency without accuracy loss. Importantly, **LLM-summarized retrieval highlights the necessity of addressing retrieval noise in RAG-RL methods**, guiding more stable and generalizable reasoning.

B.3 IMPACT OF LLM SUMMARY AND CORPUS ON UR² PERFORMANCE

Table 10 examines the robustness of UR² across different LLM summary sources. Remarkably, our framework maintains strong performance regardless of the summarization model quality. While GPT-4.1 achieves the best results (53.4% average), even using smaller open-source models like

Qwen-3-8B (52.0%) or budget-friendly APIs like GPT-4o-mini (50.8%) yields substantial improvements over Vanilla RL (48.4%). Most notably, the *w/o Summary* variant still achieves 50.2%—demonstrating that our two-stage training and retrieval-aware prompting mechanisms are inherently robust and not dependent on expensive summarization models. This flexibility makes UR² practically deployable across various computational budgets while maintaining its effectiveness, confirming the generalizability of our approach beyond specific model configurations.

Table 11 investigates the impact of different corpus configurations on UR²'s performance across open-domain QA tasks. The results reveal several key insights about corpus design choices. First, using Wikipedia abstracts (`Abs`, released with HotpotQA) versus full articles (`Full`) shows task-dependent effects: abstracts perform better on easy questions (HotpotQA), while full articles excel on complex reasoning tasks requiring broader context (2Wiki, Bamboogle, MusiQue). Second, the presence of LLM summarization consistently improves performance across all configurations, with average F1 scores increasing by 6.5–10.8% when summaries are applied. Notably, UR² maintains competitive performance even without summaries (50.6% F1 with `Abs`, 47.4% with `Full`), substantially outperforming ZeroSearch's reliance on synthetic content. The retrieval frequency (#R) analysis shows that UR² strategically balances retrieval calls—using fewer retrievals than Search-R1 while achieving superior performance, demonstrating more efficient knowledge utilization.

Table 11: Performance of UR² and baselines on open-domain QA datasets across different corpus configurations. `Abs` denotes corpora based on Wikipedia abstracts, while `Full` uses full articles. For each corpus, we use top-10 documents with summaries and top-5 without. #R represents the number of successful retrievals per question.

| Corpus | Summ. | Models | Hotpot [†] | | | 2Wiki [†] | | | Bamb. [‡] | | | MusiQ. [‡] | | | Avg | | |
|-------------------|--------------|-----------------------|---------------------|-------------|------|--------------------|-------------|------|--------------------|-------------|------|---------------------|-------------|------|-------------|-------------|------|
| | | | F1 | LSJ | #R | F1 | LSJ | #R | F1 | LSJ | #R | F1 | LSJ | #R | F1 | LSJ | #R |
| <code>Abs</code> | \checkmark | ZeroSearch | 46.0 | 50.4 | 0.66 | 38.4 | 38.6 | 0.73 | 35.8 | 38.4 | 0.54 | 14.7 | 13.8 | 0.62 | 33.7 | 35.3 | 0.64 |
| | | Search-R1 | 72.4 | 78.8 | 1.92 | 61.0 | 63.8 | 3.16 | 58.9 | 56.8 | 2.58 | 32.2 | 32.0 | 2.92 | 56.1 | 57.9 | 2.64 |
| | | R1-Searcher | 71.8 | 78.0 | 1.93 | 57.9 | 63.6 | 2.17 | 56.5 | 53.6 | 2.02 | 33.2 | 32.6 | 2.33 | 54.9 | 57.0 | 2.11 |
| | | UR² | 71.2 | 79.4 | 2.22 | 62.6 | 65.0 | 2.72 | 64.5 | 62.4 | 2.30 | 35.8 | 34.6 | 2.61 | 58.5 | 60.4 | 2.46 |
| <code>Abs</code> | \times | ZeroSearch | 44.1 | 47.0 | 0.64 | 32.9 | 31.8 | 0.66 | 32.6 | 35.2 | 0.52 | 14.3 | 11.8 | 0.61 | 31.0 | 31.5 | 0.61 |
| | | Search-R1 | 65.8 | 72.4 | 2.68 | 41.8 | 51.6 | 3.54 | 44.8 | 44.8 | 2.96 | 25.1 | 24.1 | 3.49 | 44.4 | 48.2 | 3.17 |
| | | R1-Searcher | 69.7 | 75.2 | 2.16 | 56.6 | 58 | 2.45 | 41.7 | 40.0 | 2.38 | 23.7 | 22.4 | 2.84 | 47.9 | 48.9 | 2.46 |
| | | UR² | 67.6 | 73.6 | 1.98 | 59.1 | 59.6 | 2.53 | 47.5 | 47.2 | 2.10 | 28.2 | 25.4 | 2.43 | 50.6 | 51.5 | 2.26 |
| <code>Full</code> | \checkmark | ZeroSearch | 44.3 | 48.8 | 0.74 | 36.5 | 36.8 | 0.90 | 46.3 | 44.8 | 0.70 | 19.3 | 20.0 | 0.81 | 36.6 | 37.6 | 0.79 |
| | | Search-R1 | 66.0 | 67.2 | 2.01 | 60.6 | 65.6 | 3.12 | 70.0 | 71.2 | 2.06 | 37.8 | 39.0 | 2.69 | 58.6 | 60.8 | 2.47 |
| | | R1-Searcher | 62.9 | 68.0 | 1.97 | 62.5 | 66.8 | 2.15 | 69.0 | 65.6 | 1.86 | 36.7 | 37.8 | 2.24 | 57.8 | 59.6 | 2.06 |
| | | UR² | 62.6 | 68.0 | 2.11 | 63.3 | 67.6 | 2.73 | 73.0 | 74.0 | 2.13 | 40.4 | 42.0 | 2.55 | 59.8 | 62.9 | 2.38 |
| <code>Full</code> | \times | ZeroSearch | 39.2 | 41.6 | 0.63 | 34.0 | 33.8 | 0.67 | 34.1 | 36.0 | 0.50 | 13.4 | 11.8 | 0.58 | 30.2 | 30.8 | 0.59 |
| | | Search-R1 | 57.4 | 60.6 | 2.75 | 49.2 | 51.0 | 3.50 | 57.6 | 55.2 | 2.82 | 26.9 | 26.4 | 3.40 | 47.8 | 48.3 | 3.12 |
| | | R1-Searcher | 57.6 | 61.6 | 2.24 | 56.0 | 59.0 | 2.37 | 57.5 | 57.6 | 2.07 | 26.8 | 26.6 | 2.63 | 49.5 | 51.2 | 2.33 |
| | | UR² | 54.6 | 60.6 | 2.03 | 54.5 | 55.8 | 2.51 | 52.6 | 49.6 | 2.06 | 27.8 | 26.2 | 2.38 | 47.4 | 48.1 | 2.25 |

Table 12 examines corpus selection for domain-specific tasks, comparing general Wikipedia against specialized MedQA textbooks for medical reasoning. The results demonstrate that domain-specific corpora provide marginal improvements when summarization is applied (70.2% vs. 69.6% on MedQA), but this advantage diminishes without summaries. More importantly, the performance gap between summarized and non-summarized variants is substantial (8.4% on MedQA with Wikipedia), highlighting that effective summarization is more critical than corpus specialization. This finding suggests that UR²'s LLM-summarized approach can effectively bridge the gap between general and specialized knowledge sources, making it practical for deployment across diverse domains without extensive corpus curation.

Collectively, Tables 10, 11, and 12 demonstrate UR²'s robustness across three critical dimensions: corpus configuration, domain specialization, and summarization quality. The framework maintains strong performance whether using abstracts or full articles, general or specialized corpora, and expensive or budget-friendly summarizers. Most remarkably, even without any summarization, UR² achieves competitive results through its two-stage training and difficulty-aware retrieval mech-

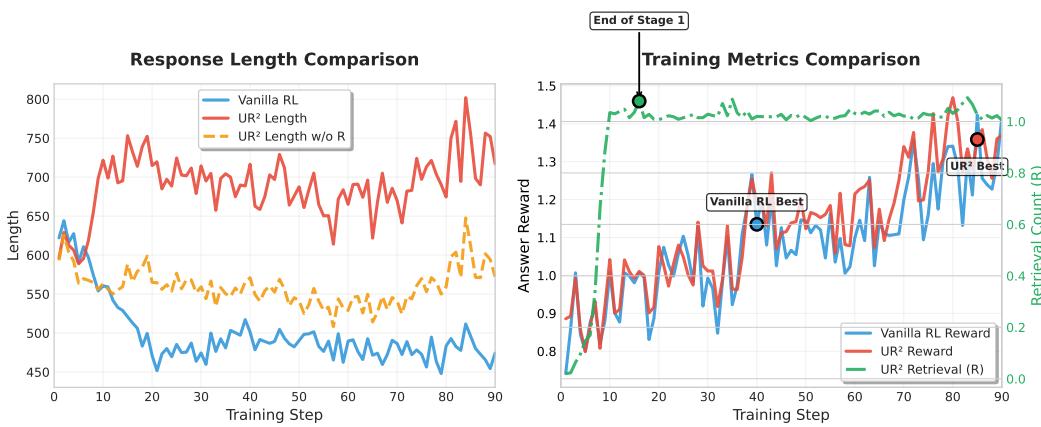
864
 865
 866
 867
 868 Table 12: Ablation study of UR² on the medical reasoning tasks. We compare different corpus
 869 (Wikipedia vs. MedQA Textbooks) and the effect of applying summarization. “w/o Summary” uses
 870 top-3 retrieved document.
 871
 872
 873

| Corpus | Summ. | Medicine [†] | M-Med [‡] |
|-----------|-------|-----------------------|--------------------|
| Wikipedia | ✓ | 69.6 | 62.8 |
| Textbooks | ✓ | 70.2 | 63.8 |
| Wikipedia | ✗ | 61.2 | 59.2 |
| Textbooks | ✗ | 62.0 | 58.0 |

874
 875
 876 This comprehensive ablation validates that UR²’s effectiveness stems from its fundamental
 877 architecture rather than dependency on specific external resources, confirming its practical applica-
 878 bility across diverse computational and domain constraints.
 879

880 B.4 COMPARATIVE ANALYSIS OF RETRIEVAL INTEGRATION IN RL TRAINING

881
 882 Figure 3 reveals key differences between UR² and Vanilla RL MCQ on Qwen-2.5-3B-Instruct in
 883 training. Vanilla RL saturates early at step 40 with 1.1 reward, with later gains mainly due to repeated
 884 data every 47 steps. In contrast, UR² steadily improves to 1.4 reward by step 85, matching the 15.7%
 885 relative benchmark gain. Retrieval frequency remains dynamic after Stage 1, showing selective use.
 886 UR² also generates longer outputs post-training, indicating deeper reasoning. This extended training
 887 capability demonstrates that retrieval-augmented approaches fundamentally expand model capacity
 888 limits, enabling continuous learning beyond traditional RL saturation points.
 889



903
 904 Figure 3: Comparison of Vanilla RL MCQ and UR² performance on Qwen-2.5-3B-Instruct across
 905 training steps. Peak test set performances are indicated.
 906

907 B.5 UNSUCCESSFUL ATTEMPTS ON REASONING MODELS

908
 909 We also conducted experiments on the R1-like model DeepSeek-R1-Distill-Qwen-7B⁵.
 910 However, when applying the MMLU-Pro prompting setup, we observed that the model lacked
 911 any retrieval capability. This remained true even after replacing the original searching special
 912 tags with alternative tokens <search></search> and <information></information>, which were shown in the ablation study (Table 4) to more effectively trigger retrieval. These results
 913 indicate a degradation of tool-usage ability after extensive chain-of-thought training. Due to compu-
 914 tational constraints, we did not extend training to more updated models such as the Qwen-3 series.
 915 We plan to supplement this work with relevant code and experiments in future updates.
 916

917
 5⁵<https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B>

918 **C TRAINING DETAILS**
 919

920 **C.1 TRAINING SETTING DETAILS**
 921

922 We train UR² using the REINFORCE++ algorithm (Guo et al., 2025a), a simplified variant of Prox-
 923 imal Policy Optimization (PPO) designed to encourage exploration. In particular, we discard the
 924 critic and omit both KL-divergence and clipping terms, following previous findings (Zhang et al.,
 925 2025; Song et al., 2025a; Chen et al., 2025) that excessive regularization can impede effective strat-
 926 egy learning in sparse-reward scenarios.

927 To reduce overfitting to retrieved content, we adopt a retrieval masking strategy (Sun et al., 2025;
 928 Song et al., 2025a; Jin et al., 2025), which treats retrieved external knowledge as part of the obser-
 929 vation space rather than trainable input. This encourages the model to reason based on retrieved
 930 information without directly optimizing on it. Our implementation builds upon the REINFORCE++
 931 baseline provided by OpenRLHF (Hu et al., 2024).

932 Each prompt is rolled out $G = 16$ times. We use the mean reward of each rollout group as the base-
 933 line for computing the advantage of each sample. To stabilize training, we apply a two-stage nor-
 934 malization scheme: normalization is first performed within each rollout group, followed by global
 935 normalization across the full batch.

936 Training is conducted with DeepSpeed ZeRO-2 (Rajbhandari et al., 2020) for memory efficiency.
 937 We use gpt-4.1-mini-2025-04-14 as the summarization model during training. Token limits
 938 per generation turn are set to 3072 for math tasks, 1536 for multiple-choice questions (MCQ), and
 939 512 for open-domain QA. Sampling parameters are fixed as temperature = 1.0 and top_p
 940 = 0.9.

941 We train for up to 2 epochs. In practice, most models achieve optimal performance within 1.5
 942 epochs. Therefore, we report results from the checkpoint with the best test set performance within
 943 the first 1.5 epochs. We save checkpoints every 5 steps for single-task training and 3B mod-
 944 els, and every 10 steps for larger-scale experiments. The specific training steps for each reported
 945 model are detailed in Table 13 below. *W/o Stage-1* variant in Table 4 replaces the special tags
 946 with <search></search> and <information></information>, removing the initial re-
 947 trieval capability activation stage. The *Weaker Stage-1* variant in Table 9 employs a modified training
 948 protocol based on UR² Qwen-2.5-7B-Instruct for MCQ tasks, where retrieval-related rewards are
 949 only provided during the initial 10 training steps. The *Qw3-8B* variant in Table 4 uses Qwen-3-8B
 950 for summarization with max_tokens = 2048, temperature = 0.3, and top_p = 0.7.
 951 Specifically, the retrieval reward assigns 0.5 for single retrieval attempts and 1.0 for multiple re-
 952 trievals (≥ 2), reflecting a more conservative retrieval activation strategy than that of our proposed
 953 method.

954 Table 13 summarizes the training configurations and checkpoint details across all model scales. Two
 955 key observations can be drawn:

956 First, for Qwen models, performance consistently improves as more training compute is introduced
 957 via our UR² method. The method’s design—encouraging structured retrieval behavior—ensures
 958 that increased steps and effective epochs lead to meaningful gains across tasks.

959 Second, while UR² also improves performance on LLaMA-3.1-8B, training on this model is ob-
 960 served to be less stable. Performance tends to saturate early (e.g., low effective epochs despite higher
 961 step counts), for both Vanilla RL and UR² variants. This indicates that LLaMA-3.1-8B may require
 962 different training strategies to maintain learning dynamics over time. Future work will explore alter-
 963 native foundation models and optimization schedules to improve convergence and stability.

964
 965 **C.2 EVALUATION DETAILS**
 966

967 All evaluations are performed using vLLM version 0.6.5. The vLLM version of Qwen-3 used
 968 is 0.8.5.post1. In evaluation, We maintain the same max_tokens limits used during train-
 969 ing: 3072 for math benchmarks, 1536 for MCQ, and 512 for open-domain QA per generation
 970 step. For GPT-family models, these limits are increased to 4096, 2048, and 1024, respectively.
 971 For sampling during evaluation, we use more conservative hyperparameters: temperature =
 0.3 and top_p = 0.5, aiming for higher answer consistency. Summarization for math tasks

972 Table 13: Training checkpoint details for UR² models. Checkpoints were saved every 5 steps for
 973 3B models and single-task training, and every 10 steps for larger models. Main experiments use the
 974 full training configuration, while ablation studies vary specific components.

| 976 Model | 977 Training Dataset | Dataset Size | Checkpoint Step | Training Epochs |
|---|----------------------|--------------|-----------------|-----------------|
| 978 Qwen-2.5-3B - Main Experiments | | | | |
| 979 UR ² -Math&QA | 980 Math&QA | 981 6000 | 982 47 | 983 1.0 |
| 984 UR ² -MCQ | 985 MMLU&Medqa | 986 9000 | 987 85 | 988 1.2 |
| 989 Vanilla RL-Math | 990 Math | 991 3000 | 992 15 | 993 0.64 |
| 994 Vanilla RL-QA | 995 QA | 996 3000 | 997 40 | 998 1.7 |
| 999 Vanilla RL-MMLU | 999 MMLU | 1000 6000 | 1001 47 | 1002 1.0 |
| 1003 Vanilla RL-MedQA | 1004 Medqa | 1005 3000 | 1006 40 | 1007 1.7 |
| 1008 Vanilla RL-MCQ | 1009 MMLU&Medqa | 1010 9000 | 1011 40 | 1012 0.57 |
| 1013 LLaMA-3.1-8B Models - Main Experiments | | | | |
| 1014 UR ² -Math&QA | 1015 Math&QA | 1016 6000 | 1017 30 | 1018 0.32 |
| 1019 UR ² -MCQ | 1020 MMLU&Medqa | 1021 9000 | 1022 30 | 1023 0.21 |
| 1024 Vanilla RL-Math | 1025 Math | 1026 3000 | 1027 30 | 1028 0.64 |
| 1029 Vanilla RL-QA | 1030 QA | 1031 3000 | 1032 47 | 1033 1.0 |
| 1034 Vanilla RL-MMLU | 1035 MMLU | 1036 6000 | 1037 60 | 1038 0.64 |
| 1039 Vanilla RL-MedQA | 1040 Medqa | 1041 3000 | 1042 30 | 1043 0.64 |
| 1044 Qwen-2.5-7B - Main Experiments | | | | |
| 1045 UR ² -Math&QA | 1046 Math&QA | 1047 6000 | 1048 40 | 1049 0.43 |
| 1050 UR ² -MCQ | 1051 MMLU&Medqa | 1052 9000 | 1053 100 | 1054 0.71 |
| 1055 Vanilla RL-Math | 1056 Math | 1057 3000 | 1058 40 | 1059 0.43 |
| 1060 Vanilla RL-QA | 1061 QA | 1062 3000 | 1063 25 | 1064 0.53 |
| 1065 Vanilla RL-MMLU | 1066 MMLU | 1067 6000 | 1068 94 | 1069 1.0 |
| 1070 Vanilla RL-MedQA | 1071 Medqa | 1072 3000 | 1073 47 | 1074 1.0 |
| 1075 Vanilla RL-MCQ | 1076 MMLU&Medqa | 1077 9000 | 1078 60 | 1079 0.43 |
| 1080 7B Models - Ablation Studies | | | | |
| 1081 Ablation-MCQ-w/o P_{fallback} | 1082 MMLU&Medqa | 1083 9000 | 1084 110 | 1085 0.78 |
| 1086 Ablation-MCQ-w/o Stage-1 | 1087 MMLU&Medqa | 1088 9000 | 1089 110 | 1090 0.78 |
| 1091 Ablation-MCQ-w/o Task Mixing | 1092 MMLU&Medqa | 1093 9000 | 1094 120 | 1095 0.85 |
| 1096 Ablation-MCQ-QW3 summary | 1097 MMLU&Medqa | 1098 9000 | 1099 50 | 1100 0.36 |
| 1101 Ablation-MCQ-4omini summary | 1102 MMLU&Medqa | 1103 9000 | 1104 50 | 1105 0.36 |
| 1106 Ablation-Math&QA weaker Stage-1 | 1107 Math&qa | 1108 6000 | 1109 80 | 1110 0.85 |
| 1111 Ablation-QA w/o LLM summary | 1112 QA | 1113 3000 | 1114 60 | 1115 1.28 |
| 1116 Ablation-QA Raw data | 1117 R1-Searcher | 1118 8148 | 1119 70 | 1120 0.55 |
| 1121 Ablation-Math Raw data | 1122 SimpleRL-Zoo | 1123 16662 | 1124 10 | 1125 0.056 |

1015
 1016
 1017
 1018 is conducted using Qwen-3-32B with `max_tokens = 8192`, `temperature = 0.3`, and
 1019 `top_p = 0.7`. Final evaluation summarization is performed using `gpt-4.1-2025-04-14`
 1020 with `max_tokens = 2048`, `temperature = 0.3`, and `top_p = 0.5`.

1021 The RL methods mentioned in this paper all follow the settings described in their original works.
 1022 Specifically, Open-Reasoner-Zero, General Reasoner, SimpleRL-Zoo, R1-Searcher, Search-R1, and
 1023 ZeroSearch are implemented using the Qwen-2.5-Base models. Although an Instruct version of
 1024 Search-R1 exists, its performance is significantly inferior and thus excluded from comparison. R1-
 1025 Searcher with LLaMA-3.1-8B adopts the Instruct variant. Vanilla methods, including CoT and
 standard RAG, are applied using the Instruct versions for all open-source models.

1026 **Advanced RAG Baseline Implementations:**

1027
 1028 **Search-o1 with Retrieval-Augmented Generation:** We adapt the Search-o1 framework (Li et al.,
 1029 2025) to operate within a controlled evaluation environment. While maintaining its core iterative
 1030 reasoning mechanism and document analysis capabilities, our implementation leverages the KILT
 1031 Wikipedia corpus with BGE-large-en-v1.5 embeddings for knowledge retrieval. This ap-
 1032 proach consolidates the multi-agent architecture into a unified model with structured prompting,
 1033 ensuring consistent evaluation across all baselines while preserving the essential reasoning patterns.
 1034

1035 **Self-Ask with Retrieval-Augmented Generation:** Our implementation follows the Self-Ask
 1036 framework’s (Press et al., 2023) question decomposition strategy, employing batch retrieval from the
 1037 local KILT corpus to enhance efficiency. The system maintains the characteristic “Follow up:” and
 1038 “Intermediate answer:” reasoning chain format, with stopping criteria incorporating both semantic
 1039 completion detection and a maximum of 10 follow-up questions. When decomposition challenges
 1040 arise, the framework seamlessly transitions to standard RAG, ensuring robust performance across
 1041 diverse question types.

1042 **RAT (Retrieval-Augmented Thought):** We adapt RAT (Wang et al., 2024b) for unified evaluation
 1043 across reasoning and QA tasks. The framework retains the core principle of knowledge-enhanced
 1044 reasoning while operating at the paragraph level rather than the sentence level, with corresponding
 1045 modifications to the prompting strategy. This design choice maintains consistency with our eval-
 1046 uation infrastructure while capturing RAT’s fundamental insight of augmenting reasoning processes
 1047 with relevant external knowledge.

1048 All advanced RAG methods operate within a standardized retrieval infrastructure: documents are
 1049 retrieved from the 100-word segmented KILT Wikipedia corpus (29M documents in total). For GPT-
 1050 family models, we use top- $k=10$ retrieval. Due to model limitations, LLaMA and Qwen variants use
 1051 top- $k=5$. For summarization or other auxiliary operations beyond reasoning, each model performs
 1052 the processing itself rather than relying on GPT-4.1, ensuring consistency with its own capabilities.

1053 **Online Corpus Retrieval Implementation:**

1054 To evaluate the generalization capability of UR² with real-world web content, we implement an
 1055 online corpus retrieval system that dynamically fetches and processes web documents. Unlike the
 1056 offline Wikipedia corpus used during training, this online retrieval mechanism provides access to
 1057 up-to-date information from the internet.

1058 The online retrieval pipeline consists of three main components:

1059 **Web Search and Content Extraction:** We utilize the **Bing Search API** to retrieve relevant URLs
 1060 based on the model’s search queries. To ensure robust retrieval quality, we implement a multi-round
 1061 crawling strategy with up to three rounds of attempts. In each round, the system fetches $k \times 3$ candi-
 1062 date URLs and crawls them in parallel using a thread pool with 256 workers. The system implements
 1063 intelligent retry logic—if the initial k URLs fail to provide sufficient valid content, it automatically
 1064 attempts to crawl additional URLs from the candidate pool. This approach significantly improves
 1065 the success rate of obtaining high-quality content.

1066 **HTML-to-Markdown Conversion:** Raw HTML content from web pages often contains noise such
 1067 as navigation elements, advertisements, and scripts. We deploy a dedicated service using ReaderLM-
 1068 v2-1.5B Model ⁶ through the vLLM framework to convert HTML to clean Markdown format. The
 1069 preprocessing pipeline removes script tags, style elements, base64-encoded images, and other irrel-
 1070 evant content using optimized regular expressions. The model then generates readable Markdown
 1071 that preserves the main textual information while discarding formatting artifacts. To improve effi-
 1072 ciency, we implement an LRU cache with a capacity of 10,000 entries, achieving significant speedup
 1073 for repeated content.

1074 **Content Summarization:** The summarization prompt is carefully designed to distinguish between
 1075 knowledge-based queries (which can be answered with factual information) and reasoning-based
 1076 queries (which require complex computation). For knowledge-based queries, the model extracts
 1077 and presents relevant facts; for reasoning-based queries, it returns a fallback message indicating that
 1078 direct reasoning is more appropriate. The summarizer here is GPT-4.1-2025-04-14.

1079 ⁶<https://huggingface.co/jinaai/reader-lm-1.5b>

1080 The entire pipeline is orchestrated through a FastAPI service that handles concurrent requests efficiently.
 1081 Rate limiting is enforced for the Bing API (95 requests per second) to comply with usage
 1082 policies. The system maintains detailed logging for debugging and performance monitoring, track-
 1083 ing metrics such as cache hit rates, crawling success rates, and end-to-end latency.

1084 Due to network and hardware limitations, a small portion of Wikipedia pages failed to be crawled
 1085 correctly, and a subset of queries did not receive valid responses. Given constraints on time and
 1086 budget, no additional remediation was applied to these cases. However, this reflects the system’s
 1087 alignment with real-world deployment settings, where large-scale QA systems must be robust to
 1088 occasional retrieval failures and operate under imperfect infrastructure.

1089 This online retrieval implementation enables UR² to access current information beyond its training
 1090 data, demonstrating its ability to integrate real-time knowledge into the reasoning process.
 1091

1092 C.3 TRAINING DATASET DETAILS

1093 We construct a unified training set that spans multiple task domains to ensure comprehensive cover-
 1094 age of diverse reasoning and knowledge-based challenges. For mathematical reasoning capabilities,
 1095 we incorporate data from the training split of SimpleZoo-RL, which provides a rich collection of
 1096 mathematical problem-solving scenarios from (Hendrycks et al., 2021; Cobbe et al., 2021). Note
 1097 that since the original SimpleZoo-RL data is relatively simple, medium- and hard-difficulty ques-
 1098 tions are largely missing, resulting in an overall easy:medium:hard ratio of 1:1:1 rather than the
 1099 7:2:1 used in Section 3.1.2. Moreover, due to limitations of LLaMA-3.1-8B-Instruct, we substitute
 1100 easy-difficulty questions for hard ones during training. To enhance open-domain QA performance,
 1101 we include samples from the R1-Searcher dataset, which spans a broad range of questions derived
 1102 from the training sets of 2Wiki and HotpotQA. For specialized domain knowledge, particularly in
 1103 the medical field, we utilize multi-choice questions from MedQA, ensuring our model can handle
 1104 domain-specific reasoning in healthcare contexts.
 1105

1106 To further diversify our training data and extend coverage to humanities subjects, we generate syn-
 1107 synthetic questions in three additional domains: philosophy, history, and economics. These synthetic
 1108 questions are created using Qwen-3-32B and follow the MMLU-Pro format to maintain consis-
 1109 tency with established academic evaluation standards. Specifically, we use 5-shot prompting with
 1110 MMLU-Pro development set examples to generate 10 questions with 4–10 options each. We discard
 1111 format-non-compliant questions and observe the model’s tendency to generate simple questions with
 1112 4–5 options, so we request the model to produce additional options and increase the difficulty for
 1113 each question. For quality control, we use GPT-4o-mini-2024-07-18 to evaluate each ques-
 1114 tion’s correctness three times, discarding any question identified as incorrect in any evaluation. We
 1115 then employ Qwen-2.5-7B-Instruct for difficulty assessment, finding approximately 80% of ques-
 1116 tions are easy-level. We randomly sample difficult questions as seeds for subsequent generations,
 1117 using different seeds for each batch. Given that downstream test sets contain subject subdivisions
 1118 (e.g., Economics encompasses microeconomics, macroeconomics, and econometrics), we utilize
 1119 Qwen-3-32B to classify questions by subdomain, ensuring comprehensive coverage. We repeat
 1120 this pipeline for 3–4 iterations to obtain the final training set.

1121 Notably, our synthetic questions differ from MMLU-Pro in emphasizing multi-hop reasoning rather
 1122 than specific knowledge points. This is evident of our results in Table 1 where Vanilla RL shows lim-
 1123 ited improvement over CoT Baseline for Qwen-2.5-7B-Instruct and LLaMA-3.1-8B-Instruct (3.9%
 1124 and –0.1% respectively), **demonstrating no overfitting to the test set**. Despite these characteris-
 1125 tics, UR² consistently achieves improvements across models, validating our method’s effectiveness.
 1126

1127 C.4 ABOUT FALBACK FAULT IN RETRIEVAL CORPUS CONSTRUCTION

1128 When the policy model generates an invalid search query that triggers a fallback message from
 1129 the LLM summarizer (i.e., *This query requires design, computation, or complex reasoning, which*
 1130 *exceeds the capabilities of a search engine. Please input another query or proceed with direct*
 1131 *reasoning.*), we observe that due to the use of retrieval masking, the model gradually learns to treat
 1132 the content within <info>...</info> as informative for reasoning. As a result, when a fallback
 1133 fault is encountered, the model tends to hallucinate. Therefore, we append the following visible
 1134 message after </info> during training to mitigate this issue: *It seems that this query exceeds*

1134 *the capabilities of the retrieval system. We may consider rephrasing it into a more fact-based and*
 1135 *searchable question that does not require complex reasoning, or proceed with direct reasoning based*
 1136 *on prior knowledge.*

1137

1138

1139 C.5 STAGE 1 TRAINING DETAILS

1140

1141 Due to the involvement of multiple models and tasks, Section 3.2.2 only presents the stage-1 setup
 1142 for Qwen-2.5-7B-Instruct on math and open-domain QA. Here, we elaborate on the initialization
 1143 strategies for other models and tasks.

1144

1145

1146 Math and Open-Domain QA. We use the discarded math training samples with rollout accuracy
 1147 below 0.2 as cold-start data. These harder examples naturally increase the likelihood of triggering
 1148 retrieval. For Qwen-2.5-3B-Instruct, its limited capacity makes it more prone to Format violations
 1149 when invoking retrieval. Since each violation incurs a -1 penalty, the original retrieval reward (+3
 1150 for one query, +4 for two or more) becomes insufficient to incentivize retrieval. To address this,
 1151 we increase the retrieval rewards to +5 and +7, respectively. In contrast, LLaMA-3.1-8B-Instruct
 1152 tends to retrieve for almost every question in early steps. To prevent over-reliance on retrieval and
 1153 preserve reasoning ability, we remove the extra reward for multiple queries and assign a fixed +3
 1154 reward upon any retrieval activation.

1155

1156

1157 MMLU-Pro and Medicine Tasks. Unlike math tasks, MMLU-Pro and medicine tasks often re-
 1158 quire domain-specific knowledge, and retrieval is less likely to lead to fallback faults. For LLaMA-
 1159 3.1-8B-Instruct and Qwen-2.5-7B-Instruct, a weak reward signal is sufficient during early training:
 1160 +0.5 for one valid retrieval and +1 for two or more. Unlike the original stage-1 design for math
 1161 and open-domain QA, this version also incorporates answer rewards from the beginning, facilitat-
 1162 ing early alignment with task-specific correctness (i.e., no longer relying on cold-start data). In this
 1163 variant, retrieval rewards are only applied during the first 10 training steps and then disabled.

1164

For Qwen-2.5-7B-Instruct trained on math and open-domain QA, we adopt the stage-1 setup origi-
 nally used for the MMLU-Pro and medicine tasks, corresponding to the weaker Stage-1 variant
 in Table 9.

1165

1166

For Qwen-2.5-3B-Instruct, we extend Stage 1 to 15 steps. To encourage retrieval, outputs that do
 not invoke any retrieval call are penalized with a -1 Format Reward (non-accumulative).

1167

1168

1169 C.6 ON RANDOMNESS AND REPRODUCIBILITY

1170

1171

RL training is known to exhibit inherent instability and variability across runs, often leading to
 divergent results even under identical settings (Nagarajan et al., 2018; Korkmaz, 2024). This ran-
 domness is attributed to factors such as stochastic policy updates, environment interactions, and
 non-deterministic hardware behavior. Despite these challenges, our experiments demonstrate re-
 markable stability. Thanks to the incorporation of Batch Normalization and Group Normalization
 in reward calculation, all models converge successfully in a single training run. The only exception
 is the UR² model of Qwen-2.5-7B on MCQ tasks, where the initial training unexpectedly resulted
 in zero retrieval activations for unknown reasons. Subsequent reruns corrected this behavior, high-
 lighting the generally robust training process in our framework.

1172

1173

During evaluation and result aggregation, we employed a non-zero temperature setting to maintain
 controlled output diversity, thereby enhancing performance and mitigating the risk of repetitive gen-
 erations. Due to the substantial API costs associated with GPT-4.1, conducting multiple evaluation
 runs to average results was not feasible. Nevertheless, given that the datasets contain approximately
 500 samples—providing sufficient statistical power—we performed a targeted reproducibility as-
 sessment on HotpotQA using the UR² Qwen 7B-Instruct model. Specifically, three independent evalua-
 tion runs yielded F1 scores of 71.7, 71.9, and 71.2, respectively. These consistent results
 indicate that stochasticity exerts minimal influence on evaluation metrics and comparative model
 assessments. Furthermore, we conducted supplementary evaluations on all identified outlier cases
 across baseline and proposed methods to ensure the robustness of our findings.

1188
1189

C.7 API CONSUMPTION

1190 We measured the API usage cost of UR² Qwen-2.5-7B-Instruct on MCQ tasks and its w/o
1191 Stage-1 variant. Training 100 steps with UR² using GPT-4.1-mini cost approximately \$320,
1192 while the w/o Stage-1 variant cost around \$100. Additionally, summarization and testing on
1193 HotpotQA using GPT-4.1 for UR² Qwen-2.5-7B-Instruct cost about \$20 per run. Since the training
1194 is a one-time expense, we consider the overall training-related consumption acceptable. Further-
1195 more, experiments reported in Section B.3 and Table 4 show that substantial performance gains can
1196 be achieved without relying on closed-source models, suggesting that open-source models or less
1197 expensive APIs provide a viable alternative for achieving comparable improvements.
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241

1242 D PROMPTS USED IN EXPERIMENTS
12431244 D.1 PROMPTS OF LLM-AS-A-JUDGE
12451246 Prompt for Math Evaluation
12471248 **Instruction:**

1249 You are an expert math evaluator. Given a question, a gold answer and a predicted answer, judge if
1250 they are mathematically consistent.

1251 Ignore formatting (e.g., \text{}, spacing, capitalization). Accept equivalent expressions (e.g., fac-
1252 torized vs expanded form). If the prediction matches only part of a multi-part answer (e.g., one of several
1253 intervals or roots), label it as **Partially correct**.

1254 **Output format:**

- 1255 • Reason: Brief explanation
- 1256 • Judgment: Correct / Partially correct / Incorrect

1257 **Input:**

- 1258 • Question: {question}
- 1259 • Gold: {gold}
- 1260 • Pred: {pred}

1263 Prompt for RAG Evaluation
12641265 **Instruction:**

1266 Given a Question and its Golden Answer, verify whether the Predicted Answer is correct. The predic-
1267 tion is correct if it fully aligns with the meaning and key information of the Golden Answer. Respond
1268 with True if the prediction is correct and False otherwise.

1269 **Input:**

- 1270 • Question: {question}
- 1271 • Golden Answer: {gold_answer}
- 1272 • Predicted Answer: {predicted_answer}

1273 **Your response should be exactly "True" or "False"**
1274

1296
1297

D.2 PROMPTS OF EVALUATION AND TRAINING

1298
1299

Prompt for MMLU-Pro&MedQA

Instruction:

You are solving a multiple-choice question. Analyze each option carefully and logically. Think step by step: consider the meaning and implications of each option, eliminate incorrect ones with clear reasoning, and select the best answer through comparison.

During your reasoning, if you're unsure about any fact, you may issue a **search query** like this: <|begin_of_query|> your concise query (less than 20 words) <|end_of_query|>

- You can issue **multiple queries** at different steps in your reasoning.
- **Each query must target only one fact or statement.** Do not combine multiple ideas in a single query.
- **Examples:**
 - ✓ <|begin_of_query|> What are the common symptoms of pneumonia?
<|end_of_query|>
 - ✓ <|begin_of_query|> What is the typical treatment for pneumonia in elderly patients? <|end_of_query|>
 - ✗ <|begin_of_query|> What are the symptoms and treatments for pneumonia in elderly patients? <|end_of_query|>
- You may issue **at most four queries** in total — use them wisely.

Once documents are returned in this format:

<|begin_of_documents|> ... (search results here) <|end_of_documents|>

Use the retrieved documents to verify, reject, or revise your prior reasoning about the options. Then continue analyzing the options until you're confident in your answer.

Final answer format: the correct answer is: A, B, C, D, etc. (only the letter corresponding to the correct option)

1325

Prompt for Math

Instruction:

You are solving a math problem. Think step by step to solve it.

The reasoning process includes detailed considerations such as analyzing questions, summarizing relevant findings, brainstorming new ideas, verifying the accuracy of current steps, refining any errors, and revisiting previous steps.

During your reasoning, if you're unsure about a factual concept — such as a definition, formula, theorem, or mathematical constant — you may issue a **search query** to clarify it.

Format your query using the following template (each query must target only one fact):

<|begin_of_query|> your concise query (less than 20 words) <|end_of_query|>

✓ Examples:

- <|begin_of_query|> Definition of Möbius function <|end_of_query|>
- <|begin_of_query|> Formula for variance of Bernoulli distribution
<|end_of_query|>

✗ Do NOT query for reasoning-related content like:

- Whether a solution approach is valid
- How to compute a specific value
- Multi-step deductions or conclusions

You may issue at most **four** search queries per problem — use them wisely.

When documents are returned in this format:

<|begin_of_documents|> ... (search results here) <|end_of_documents|>

Use the evidence to confirm or revise your reasoning. Then continue analyzing the question until you're confident in the answer.

At the end of your reasoning, give your final answer in the following format:

\boxed{YOUR_ANSWER}

1350 **Prompt for Open-Domain QA**

1351

1352 **Instruction:**

1353 You are solving a factual open-domain question from a Knowledge Question Answering (KQA) task.

1354 The question requires step-by-step reasoning over real-world knowledge to identify a specific, factually

1355 correct answer.

1356 Carefully analyze the question to understand the key entities, relationships, and constraints involved.

1357 Retrieve and consider relevant factual knowledge, and reason logically to identify the most accurate

1358 answer.

1359 During your reasoning, if you're unsure about any fact, you may issue a **search query** like this:

1360 <|begin_of_query|> your concise query (less than 20 words) <|end_of_query|>

1361 • You can issue **multiple queries** at different steps in your reasoning.

1362 • **Each query must target only one fact or statement.** Do not combine multiple ideas in a

1363 single query.

1364 – ✓ **Example:**

1365 * <|begin_of_query|> When did Einstein move to the United States?

1366 * <|begin_of_query|> Why did Einstein leave Germany?

1367 * <|begin_of_query|>

1368 – ✗ **Do not combine them like this:**

1369 * <|begin_of_query|> When did Einstein move to the US and why did he leave

1370 Germany? <|end_of_query|>

1371 • You may issue **at most five queries** in total — use them wisely.

1372 Once documents are returned in this format:

1373 <|begin_of_documents|> ... (search results here) <|end_of_documents|>

1374 Use the evidence to confirm or revise your reasoning. Then continue analyzing the question until you're

1375 confident in the answer.

1376 At the end of your reasoning, give your final answer in the following format:

1377 \boxed{YOUR_ANSWER}

D.3 PROMPTS FOR SUMMARIZING

1379

1380 **Prompt for Summarizing Math Documents During Evaluation**

1381

1382

1383 **Task Instruction:**

1384 You are assisting in solving a math problem. You are tasked with reading and analyzing Wikipedia

1385 content based on the following inputs: **Previous Reasoning Steps**, **Current Search Query**, and

1386 **Wikipedia Content**. Your task is to extract accurate and relevant information from the provided

1387 Wikipedia content to support or enhance the reasoning process.

1388 • Carefully read the provided **Wikipedia Content**;

1389 • Extract factual information that can:

1390 – Directly assist in answering the **Current Search Query**, or

1391 – Help validate, complete, or correct earlier reasoning steps.

1392 • The extracted information should be:

1393 – Accurate and trustworthy;

1394 – Closely relevant to the query;

1395 – Helpful in improving, expanding, or supporting the mathematical reasoning.

1396

1397 Important: Do NOT attempt to correct or rewrite the previous reasoning. Treat it only as contextual

1398 reference that may be flawed.

1399 **Output Format:**

1400 Present the information beginning with the label ****Final Information****

1401 as shown below.

1402 ****Final Information****
 [Helpful factual information]

1403 **Inputs:**

- 1404
 1405 • Previous Reasoning Steps: {prev_reasoning}
 1406 • Current Search Query: {search_query}
 1407 • Wikipedia Content: {wikipedia_content}

Prompt for Summarizing Math Documents During Training

Task Instruction:

You are assisting in solving a math problem. Your task is to determine whether the current query requires external factual knowledge (such as definitions, formulas, theorems, or lookup values), and if so, extract accurate and relevant information from the provided Wikipedia content to support or enhance the reasoning process.

Step 1: Classify the Query Type

Determine whether the query falls into one of the following categories:

- **Knowledge-based query:** Can be directly answered using factual knowledge.
- **Reasoning-based query:** Requires multi-step deduction, logical reasoning, or constructive computation.

If reasoning-based, return: *This query requires design, computation, or complex reasoning, which exceeds the capabilities of a search engine. Please input another query or proceed with direct reasoning.*

Step 2: Analyze Knowledge-Based Queries (if applicable)

- Carefully read the Wikipedia Content;
- Extract factual information that:
 - Directly assists the query, or
 - Helps validate, complete, or correct earlier reasoning.
- Ensure information is accurate, relevant, and objective.

Do NOT attempt to correct prior reasoning. Treat it as possibly flawed context.

Output Format:

Final Information

[Helpful factual information, or the non-knowledge-based response]

Inputs:

- Previous Reasoning Steps: {prev_reasoning}
- Current Search Query: {search_query}
- Wikipedia Content: {wikipedia_content}

Prompt for Summarizing Other Documents During Evaluation

Task Instruction:

You are tasked with reading and analyzing Wikipedia content based on the following inputs: **Previous Reasoning Steps**, **Current Search Query**, and **Wikipedia Content**. Your objective is to extract factual and relevant information from the **Wikipedia Content** that directly supports or informs the **Current Search Query**, and integrate it into the reasoning process in an objective and helpful manner.

Guidelines:

- Analyze Wikipedia Content:
 - Read carefully.
 - Identify factual info directly related to the query.
- Maintain Objectivity:
 - Do not validate or revise prior reasoning.
 - Use it as flawed context.

Output Format:

Final Information

[Helpful information]

1458

1459

Inputs:

- Previous Reasoning Steps: {prev_reasoning}
- Current Search Query: {search_query}
- Wikipedia Content: {wikipedia_content}

1464

1465

Prompt for Summarizing Other Documents During Training

1466

1467

Task Instruction:

Your first task is to determine whether the provided query is a knowledge-based query that can be answered using factual information from Wikipedia, or if it requires design, computation, or complex reasoning.

1471

Step 1: Query Classification

1472

1473

- If knowledge-based (e.g., facts, definitions, history), proceed to Step 2.
- Otherwise, return:

1474

1475

This query requires design, computation, or complex reasoning, which exceeds the capabilities of a search engine. Please input another query or proceed with direct reasoning.

1476

Step 2: Analyze Knowledge-Based Queries

1477

1478

1479

1480

- Read Wikipedia content;
- Extract relevant factual information;
- Stay neutral—do not alter previous reasoning;

1481

Output Format:

1482

Final Information

1483

[Helpful information or the non-knowledge-based response]

1484

Inputs:

1485

1486

1487

1488

- Previous Reasoning Steps: {prev_reasoning}
- Current Search Query: {search_query}
- Wikipedia Content: {wikipedia_content}

1489

1490

1491

1492

1493

1494

1495

1496

1497

1498

1499

1500

1501

1502

1503

1504

1505

1506

1507

1508

1509

1510

1511

1512 D.4 PROMPTS FOR BASELINE METHODS
15131514 **Self-Ask Initial Prompt**1515 **Instruction:**

1516 The self-ask method uses few-shot examples to demonstrate the reasoning pattern:
 1517

1518 **Example 1:**

1519 Question: Who lived longer, Muhammad Ali or Alan Turing?
 1520 Are follow up questions needed here: Yes.
 1521 Follow up: How old was Muhammad Ali when he died?
 1522 Intermediate answer: Muhammad Ali was 74 years old when he died.
 1523 Follow up: How old was Alan Turing when he died?
 1524 Intermediate answer: Alan Turing was 41 years old when he died.
 1525 So the final answer is: Muhammad Ali

1526 **Example 2:**

1527 Question: When was the founder of craigslist born?
 1528 Are follow up questions needed here: Yes.
 1529 Follow up: Who was the founder of craigslist?
 1530 Intermediate answer: Craigslist was founded by Craig Newmark.
 1531 Follow up: When was Craig Newmark born?
 1532 Intermediate answer: Craig Newmark was born on December 6, 1952.
 1533 So the final answer is: December 6, 1952

1534 **Example 3:**

1535 Question: Who was the maternal grandfather of George Washington?
 1536 Are follow up questions needed here: Yes.
 1537 Follow up: Who was the mother of George Washington?
 1538 Intermediate answer: The mother of George Washington was Mary Ball Washington.
 1539 Follow up: Who was the father of Mary Ball Washington?
 1540 Intermediate answer: The father of Mary Ball Washington was Joseph Ball.
 1541 So the final answer is: Joseph Ball

1542 **Input:**

1543 Question: {question}
 1544 Options: {options}
 1545 Are follow up questions needed here:
 1546

1547 **Self-Ask Sub-question Answering Prompt**1548 **Instruction:**

1549 Please answer the following question based on the reference text. If the reference text does not contain
 1550 sufficient information to answer the question, you may use your own knowledge to provide the answer.
 1551 Always think step by step.
 1552 Provide your final answer in the format \boxed{YOUR_ANSWER}.

1553 **Input:**

- 1554 • Question: {subquestion}
- 1555 • Reference text: {reference}

1566

RAT Draft Generation Prompt

1567

1568

System Prompt:

1569

You are an advanced AI assistant tasked with answering open-domain questions. You excel at providing comprehensive, well-structured answers with multiple paragraphs. Each paragraph you write contains multiple sentences that thoroughly explore the topic. You always follow formatting instructions precisely.

1570

Instruction:

1571

IMPORTANT: Structure your response as follows:

1572

1. Write a comprehensive answer with MULTIPLE PARAGRAPHS (3-6 paragraphs typically).

1573

2. Each paragraph MUST contain AT LEAST 2 complete sentences. Single-sentence paragraphs are NOT acceptable.

1574

3. Separate paragraphs with blank lines (press Enter twice).

1575

4. At the very end, after all paragraphs, add your final answer in this format:

1576

\box{ANSWER}

1577

where ANSWER is ONLY the direct answer - typically just a name, number, date, or short phrase.

1578

Examples:

1579

- For “Who was the first president?” → \box{George Washington}

1580

- For “When was the company founded?” → \box{1812}

1581

- For “What is the capital?” → \box{Paris}

1582

DO NOT include explanations or full sentences in the box.

1583

Input:

1584

- Question: {question}

1585

1586

1587

1588

1589

RAT Query Generation Prompt

1590

1591

Instruction:

1592

Based on the question and the current answer content, generate a search query to verify or find additional information.

1593

Please summarize the content with the corresponding question. This summarization will be used as a query to search with Bing search engine. The query should be short but need to be specific to promise Bing can find related knowledge or pages. You can also use search syntax to make the query short and clear enough for the search engine to find relevant language data. Try to make the query as relevant as possible to the last few sentences in the content.

1594

IMPORTANT: Just output the query directly. DO NOT add additional explanations or introducement in the answer unless you are asked to.

1595

Input:

1596

1597

1598

1599

1600

- Question: {question}

1601

1602

1603

1604

1605

1606

1607

1608

1609

1610

1611

1612

1613

1614

1615

1616

1617

1618

1619

1620

RAT Answer Revision Prompt

1621

1622

Instruction:

1623

I want to revise the answer according to retrieved related text of the question. You need to check whether the answer is correct. If you find some errors in the answer, revise the answer to make it better. If you find some necessary details are ignored, add it to make the answer more plausible according to the related text.

1624

IMPORTANT:

1625

1626

1627

1628

1629

1630

1631

1632

1633

1634

1. Keep the structure with multiple substantial paragraphs.
2. Use blank lines to separate paragraphs (press Enter twice).
3. If the original answer has `\boxed{...}` at the end, you MUST keep it and update it if needed.
4. The `\boxed{}` should contain ONLY the direct answer (name/number/date/short phrase), NOT a full sentence.

Just output the revised paragraphs directly, including the `\boxed{}` if present.

Input:

1635

1636

1637

1638

1639

1640

1641

Search-o1 Reasoning Prompt

1642

System Prompt:

1643

You are a reasoning assistant with the ability to perform web searches to help you answer the user's question accurately. You have special tools:

1644

1645

1646

1647

1648

1649

1650

- To perform a search: write `<|begin_search_query|>` your query here `<|end_search_query|>`.
- Then, the system will search and analyze relevant web pages, then provide you with helpful information in the format `<|begin_search_result|>` ...search results... `<|end_search_result|>`.

You can repeat the search process multiple times if necessary. The maximum number of search attempts is limited to `{max_rounds}`.

Once you have all the information you need, continue your reasoning.

Example:

Question: "Alice David is the voice of Lara Croft in a video game developed by which company?"

Assistant thinking steps:

1651

1652

1653

1654

1655

1656

1657

1658

1659

1660

1661

1662

1663

1664

1665

1666

1667

1668

1669

1670

1671

1672

1673

1674
1675**Prompt for MMLU-Pro&MedQA (CoT)**

1676

Instruction:1677
1678

You are solving a multiple-choice question. Think step by step and use careful reasoning. For each question, **analyze all options one by one**. For each option:

1679

- Consider its meaning and implications.
- Evaluate whether it is correct or incorrect, and **explain why**.
- Eliminate incorrect options with clear, logical reasoning.

1683

After analyzing all options, compare the remaining ones and choose the best answer.

1684
1685

At the end of your reasoning, give your final answer in the following format:
the correct answer is: A, B, C, D, etc. (only the letter corresponding to the correct option).

1686

Input:1687
1688

- Question: {question}
- Options: {options}

1690
1691
1692**Prompt for Math (CoT)**1693
1694**Instruction:**1695
1696

Please answer the following math question. You should think step by step to solve it.
Provide your final answer in the format \boxed{YOUR_ANSWER}.

Input:1697
1698

- Question: {question}

1699
1700
1701**Prompt for Open-Domain QA (CoT)**1702
1703
1704**Instruction:**1705
1706

{question}

Please reason step by step, and put your final answer within \boxed{}.

1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727

1728 **E CASE STUDY**
 1729

1730 **E.1 SEARCH INDEPENDENTLY BASED ON DIFFERENT DIFFICULTY LEVELS AND QUESTIONS**
 1731

1732 **Simple questions are not searched**

1733 **Question:**

1734 Let $z = 2 + \sqrt{2} - (3 + 3\sqrt{2})i$, and let $c = 2 - 3i$. Let w be
 1735 the result when z is rotated around c by $\frac{\pi}{4}$ counter-clockwise.
 1736 **Ground Truth:** $6 - 5i$

1737 **Reasoning:**

1738 To find the complex number w after rotating z around c by $\frac{\pi}{4}$
 1739 counter-clockwise, we will follow these steps:
 1740

1741 1. Express z and c in complex form:
 1742

$$z = 2 + \sqrt{2} - (3 + 3\sqrt{2})i$$

$$c = 2 - 3i$$

1743 2. Translate z and c so that c is at the origin:
 1744

$$z' = z - c = (2 + \sqrt{2} - (3 + 3\sqrt{2})i) - (2 - 3i) = \sqrt{2} - 3\sqrt{2}i$$

1745 3. Rotate z' by $\frac{\pi}{4}$ counter-clockwise:
 1746

1747 The rotation of a complex number z' by θ counter-clockwise is given by multiplying z' by $e^{i\theta}$. For $\theta = \frac{\pi}{4}$:

$$e^{i\frac{\pi}{4}} = \cos\left(\frac{\pi}{4}\right) + i \sin\left(\frac{\pi}{4}\right) = \frac{\sqrt{2}}{2} + i\frac{\sqrt{2}}{2}$$

1748 Therefore,

$$z'' = z' \cdot e^{i\frac{\pi}{4}} = (\sqrt{2} - 3\sqrt{2}i) \left(\frac{\sqrt{2}}{2} + i\frac{\sqrt{2}}{2} \right)$$

1749 4. Perform the multiplication:
 1750

$$z'' = \sqrt{2} \cdot \frac{\sqrt{2}}{2} + \sqrt{2} \cdot i \frac{\sqrt{2}}{2} - 3\sqrt{2}i \cdot \frac{\sqrt{2}}{2} - 3\sqrt{2}i \cdot i \frac{\sqrt{2}}{2}$$

1751 Simplify each term:
 1752

$$z'' = \frac{2}{2} + i\frac{2}{2} - i\frac{6}{2} + \frac{6}{2}$$

$$z'' = 1 + i - 3i + 3$$

1753 Combine like terms:
 1754

No search

$$z'' = 4 - 2i$$

1755 5. Translate z'' back by adding c :
 1756

$$w = z'' + c = (4 - 2i) + (2 - 3i) = 6 - 5i$$

1757 Thus, the final answer is:
 1758

$$6 - 5i$$

1759 **Figure 4: UR² can search independently based on different difficulty levels and questions.**
 1760 **Left:** For simpler questions, UR² considers that external knowledge support is not necessary and
 1761 does not perform a search. **Right:** For questions of a certain difficulty, UR² will actively search
 1762 for answers.

1763 **Difficult questions are searched**

1764 **Question:**

1765 In the state of Coinland, coins have values 6, 10, and 15 cents.
 1766 Suppose x is the value in cents of the most expensive item in
 1767 Coinland that cannot be purchased using these coins with exact
 1768 change. What is the sum of the digits of x ?
 1769

1770 **Ground Truth:** 11.0

1771 **Reasoning:**

1772 **To solve this problem, we need to find the largest value that
 1773 cannot be expressed as a combination of the coin values 6,
 1774 10, and 15 cents.**

1775 **<se>**
 1776 **Definition of the Frobenius coin problem for three coin val-**
 1777 **ues.**
 1778 **</se>**
 1779 **<info>**

1780 **For three coin denominations, the Frobenius coin problem
 1781 seeks the largest monetary amount that cannot be obtained
 1782 using any combination of the given denominations.**

1783 **... (more search information content)**

1784 **In the case of denominations 6, 10, and 15, since their GCD
 1785 is 1, the Frobenius number is guaranteed to exist, but its
 1786 exact value must be computed algorithmically or via spe-**

1787 **cialized methods for three variables.**

1788 **</info>**

1789 **Search**

1790 Since there is no simple formula for three coin values, we
 1791 will use a computational approach or known results to find the
 1792 largest value that cannot be expressed as a combination of 6,
 1793 10, and 15.

1794 By checking manually or using a computational tool, we find
 1795 that the largest value that cannot be expressed as a combination
 1796 of 6, 10, and 15 is **29 cents**.

1797 To confirm, we can check if 29 can be expressed as a combi-
 1798 nation of 6, 10, and 15:

- $29 - 15 = 14$ (not a combination of 6 and 10)
- $29 - 10 = 19$ (not a combination of 6 and 10)
- ... (more computational validation)

1799 Since 29 cannot be expressed as a combination of 6, 10, and
 1800 15, we check the next value, 30, which can be expressed as:

$$30 = 2 \times 15$$

1801 Thus, the largest value that cannot be purchased using these
 1802 coins is:

$$29$$

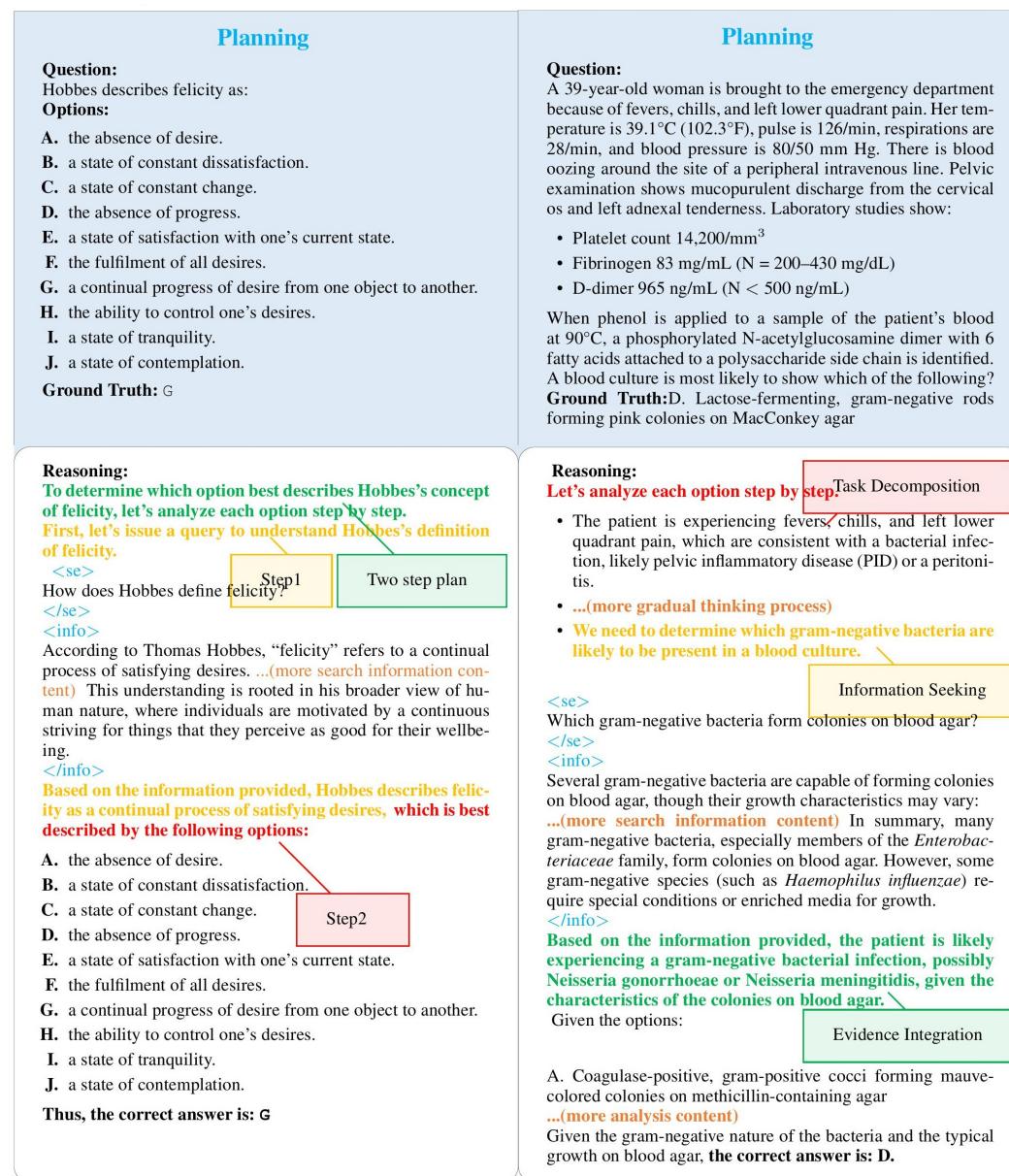
1803 The sum of the digits of 29 is:

$$2 + 9 = 11$$

1782 **E.2 PLANNING**

1783

1784



1823 Figure 5: UR² can formulate plans and dynamically adjust them during the reasoning process.
Left: UR² develops a plan and completes it in two steps. The first step is to search for the required knowledge, and the second step is to check each option individually. **Right:** UR² demonstrates clear planning behavior by decomposing the diagnostic task into sequential reasoning steps, identifying knowledge gaps, and querying external information to support its final decision.

1824

1825

1826

1827

1828

1829

1830

1831

1832

1833

1834

1835

1836 E.3 CROSS VALIDATION
 1837

1838

1839

1840

1841

1842

1843

1844

1845

1846

1847

1848

1849

1850

1851

1852

1853

1854

1855

1856

1857

1858

1859

1860

1861

1862

1863

1864

1865

1866

1867

1868

1869

1870

1871

Cross Validation

Question:
What is the name of the 1999 film directed by Roman Polanski and starring his wife Emmanuelle Seigner that was a world-wide success but critically and commercially unsuccessful in North America?

Ground Truth: The Ninth Gate

Reasoning:
To answer this question, I need to identify the 1999 film directed by Roman Polanski that starred his wife Emmanuelle Seigner, and then determine if it was a worldwide success but critically and commercially unsuccessful in North America.

<se>
Who starred in films directed by Roman Polanski in 1999?
</se>
<info>

According to the Wikipedia content, Roman Polanski directed the film "The Ninth Gate" in 1999. The content does not provide a full cast list, but it confirms that "The Ninth Gate" is a 1999 mystery thriller film directed by Roman Polanski. ...**(more search information content)** but it does confirm that "The Ninth Gate" is the relevant 1999 film directed by Roman Polanski.
</info>

Since "The Ninth Gate" was the 1999 film directed by Roman Polanski, I need to check if Emmanuelle Seigner starred in it and if it had the described success pattern.

<se>
Did Emmanuelle Seigner star in "The Ninth Gate"?
</se>
<info>

According to the Wikipedia content, Emmanuelle Seigner is known for her role in "The Ninth Gate" (1999). This confirms that Emmanuelle Seigner starred in "The Ninth Gate," the 1999 film directed by Roman Polanski.
</info>

Now that we know "The Ninth Gate" was the 1999 film directed by Roman Polanski and starring Emmanuelle Seigner, I need to verify if it was a worldwide success but critically and commercially unsuccessful in North America.

<se>
Was "The Ninth Gate" a worldwide success but critically and commercially unsuccessful in North America?
</se>
<info>

According to the Wikipedia content ...**(more search information content)** However, despite its poor reception in North America, the film earned a worldwide gross of \$58.4 million against a \$38 million budget, indicating that it achieved commercial success internationally.
</info>

Based on the evidence, "The Ninth Gate" was indeed a worldwide success but critically and commercially unsuccessful in North America.

Thus, the correct answer is: The Ninth Gate

Figure 6: This case demonstrates UR² cross validation ability by verifying the same candidate answer "The Ninth Gate" from multiple angles. It confirms the director and year, validates the actress's participation, and contrasts international vs. North American commercial and critical success. The multi-step confirmation strengthens the confidence in the final answer, showcasing robust cross validation behavior.

1872

1873

1874

1875

1876

1877

1878

1879

1880

1881

1882

1883

1884

1885

1886

1887

1888

1889

1890 **E.4 REFLECTION(TRY RETRIEVE/REFINE QUERY)**

1891

1892 **Reflection (Retry Retrieve/Refine Query)**

1893 **Question:**

1894 A 3-year-old boy is brought to the physician for the evaluation of recurrent skin lesions. The episodes of lesions started at the age of
 1895 2 months and multiple treatment options have been attempted without success. He has also had several episodes of respiratory tract
 1896 infections, enlarged lymph nodes, and recurrent fevers since birth. The boy attends daycare. His older brother has asthma. The patient's
 1897 immunizations are up-to-date. He is at the 5th percentile for length and 10th percentile for weight. He appears ill. His temperature
 1898 is 38°C (100.4°F), pulse is 100/min, and blood pressure is 100/60 mm Hg. Examination shows several raised, erythematous lesions
 1899 of different sizes over the face, neck, groins, and extremities; some secrete pus. Cervical and axillary lymph nodes are enlarged
 bilaterally. The remainder of the examination shows no abnormalities. Which of the following is the most likely diagnosis?

1900 **Options:**

- A. Chronic granulomatous disease
- B. Ataxia-telangiectasia
- C. Atopic dermatitis
- D. Wiskott-Aldrich syndrome
- E. Chediak-Higashi syndrome

1901 **Ground Truth:** A

1902 **Reasoning:**

1903 Let's analyze each option step by step.

- The patient has recurrent skin lesions, respiratory tract infections, enlarged lymph nodes, and recurrent fevers since birth.
- The patient is at the 5th percentile for length and 10th percentile for weight, appearing ill.
- The temperature is 38°C (100.4°F), pulse is 100/min, and blood pressure is 100/60 mm Hg.
- Examination shows raised, erythematous lesions of different sizes over the face, neck, groins, and extremities, some secrete pus. Cervical and axillary lymph nodes are enlarged bilaterally.

1904 **Let's issue a query to determine the likely diagnosis.**

1905 <se>

1906 **Which disease causes recurrent skin lesions and respiratory infections in a child?**

1907 </se>

1908 <info>

1909 Several rare immunodeficiency disorders can cause recurrent skin lesions and respiratory infections in children:

- **DOCK8 deficiency**: This is a combined immunodeficiency syndrome characterized by recurrent skin infections (bacterial, viral, and fungal), severe eczema, and recurrent respiratory tract infections such as sinusitis, otitis media, pneumonia, and bronchitis. Patients may also have recurrent fevers, lymphadenopathy (enlarged lymph nodes), and are prone to both skin and pulmonary infections.

1910 ...**(more search information content)**

1911 These conditions are characterized by a combination of recurrent skin and respiratory infections, often accompanied by other signs of
 1912 immune dysfunction such as lymphadenopathy, eczema, and growth failure. Among these, DOCK8 deficiency is particularly notable
 1913 for the combination of severe skin infections, eczema, and recurrent respiratory tract infections in children.

1914 </info>

1915 **Let's issue another query to narrow down the options.**

1916 <se>

1917 **Which immunodeficiency causes raised, erythematous lesions with pus?**

1918 </se>

1919 <info>

1920 Among primary immunodeficiencies, several can present with recurrent skin infections and lesions. According to the Wikipedia
 1921 content, primary immunodeficiency diseases often lead to increased susceptibility to recurrent skin infections, and some, such as
 1922 Hyperimmunoglobulin E syndrome (also known as Job's syndrome), are specifically noted for causing skin lesions. ...**(more search**

1923 **information content)** erythematous, pus-filled skin lesions.

1924 Therefore, Hyperimmunoglobulin E syndrome (Job's syndrome) is the immunodeficiency most characteristically associated with
 1925 raised, erythematous skin lesions that may secrete pus.

1926 </info>

1927 Based on the information provided, the patient's symptoms are consistent with Hyperimmunoglobulin E syndrome (Job's syndrome),
 1928 which is characterized by recurrent skin infections with pus-secreting lesions.

1929 **Thus, the correct answer is: A**

1930 Refine Query

1931 Figure 7: After an initial analysis of the patient's symptoms, UR² conducted two rounds of retrieval:
 1932 first a broad query ("Which disease causes..."), then refined to a targeted question ("Which immunodeficiency causes...") after recognizing the initial query was insufficient. This fully demonstrates
 1933 the LLM's capability to refine queries.

1934

1935

1936

1937

1938

1939

1940

1941

1942

1943

1944
 1945
 1946
 1947
 1948
 1949
 1950 **Reflection (Retry Retrieve/Refine Query)**
 1951 **Question:**
 1952 Marcia Baron argues that the artificiality of ticking bomb cases is problematic because
Options:
 1953 A. the artificial assumptions are, in fact, empirical claims that are central to the debate.
 1954 B. it leads us to believe that torture is always permissible.
 1955 C. all artificial thought experiments are philosophically useless.
 1956 **Ground Truth:** A
 1957
 1958 **Reasoning:**
 1959 To determine which option correctly describes Marcia Baron's argument about the artificiality of ticking bomb cases, let's analyze
 1960 the statement and each option step by step.
 1961 **First, let's search for the specific reason Marcia Baron gives for why the artificiality of ticking bomb cases is problematic.**
 1962 <se>
 1963 Why does Marcia Baron argue that ticking bomb cases are problematic?
 1964 <se>
 1965 <info>
 1966 The Wikipedia content provided does not include a direct summary of Marcia Baron's specific arguments about the problematic nature
 1967 of ticking bomb cases. However...(more search information content) These hidden and often implausible assumptions obscure the
 1968 true costs and complexities of permitting torture in real-life situations. While Marcia Baron's own detailed reasoning is not quoted in
 1969 the provided content, it is common among philosophers critical of the ticking bomb scenario to highlight its artificiality and the way
 1970 it distorts moral and practical considerations about torture.
 1971 </info>
 1972 **Based on the information provided, it seems that Marcia Baron's argument likely focuses on the unrealistic assumptions**
 1973 **inherent in ticking bomb cases. Let's compare this with the options given in the question.**
 1974 **Option A:**
 1975 "*the artificial assumptions are, in fact, empirical claims that are central to the debate.*"
 1976 This option seems to suggest that the artificial assumptions are actually empirical claims, which is not a common interpretation of
 1977 Marcia Baron's argument. The argument is more about the unrealistic nature of the assumptions rather than their empirical status.
 1978 **Option B:**
 1979 "*it leads us to believe that torture is always permissible.*"
 1980 This option seems to suggest that the artificiality leads to a false belief about the permissibility of torture, which is not a common
 1981 interpretation of Marcia Baron's argument. The argument is more about the unrealistic assumptions rather than the permissibility of
 1982 torture.
 1983 **Option C:**
 1984 "*all artificial thought experiments are philosophically useless.*"
 1985 This option seems to suggest that all artificial thought experiments are useless, which is not a common interpretation of Marcia
 1986 Baron's argument. The argument is more about the specific unrealistic assumptions in ticking bomb cases rather than the usefulness
 1987 of all artificial thought experiments.
 1988 **From the information provided, it seems that none of the options fit perfectly, but Option A might be the closest if we interpret**
 1989 **"empirical claims" as a way to describe the assumptions, even though it's not the primary focus.**
 1990 **Let's refine with a more specific search:**
 1991 <se>
 1992 What specific unrealistic assumptions does Marcia Baron criticize in ticking bomb cases? Refined Query
 1993 </se>
 1994 <info>
 1995 The Wikipedia content does not provide a direct quotation or detailed summary of Marcia Baron's own words regarding the specific
 1996 unrealistic assumptions she criticizes in ticking bomb cases. However, it does outline the general criticisms made by opponents of
 1997 the ticking time bomb scenario—criticisms that are commonly attributed to philosophers like Baron. ...**(more search information**
 1998 **content)** While Marcia Baron's specific phrasing is not included in the Wikipedia content, her critique is generally understood to
 1999 align with these points about the unrealistic and misleading nature of the assumptions underlying ticking bomb cases.
 2000 </info>
 2001 Based on the detailed information provided, Option A seems to be the best fit, even though it might be a bit stretched:
 2002 **Thus, the correct answer is: A**

Figure 8: In this case, UR² initially retrieved general critiques related to ticking bomb scenarios but found no direct reference to Marcia Baron's argument. Recognizing this gap, it engaged in Reflection by refining the query to focus on "specific unrealistic assumptions", thereby exemplifying its capability to self-correct and deepen its information retrieval process.

1998
 1999
 2000
 2001
 2002
 2003
Reflection (Retry Retrieve/Refine Query)
 2004
Question:
 This question refers to the following information.
Source 1:
 "The Sovereign Congress of Venezuela, to which authority the people of the Republic of the New Granada have voluntarily stood by.
 Considering:
 1. That united in a single Republic, the provinces of Venezuela and the New Granada have all proportions and ways to elevate themselves to a higher grade of power and prosperity.
 ...
Source 2:
 ...
Question: Which of the following would most undermine the expectations expressed in Article 1 of Source 1?
Options:
 A. Intervention by the United States and European powers in the independence of Latin American countries
 B. Nationalism in Latin America pressing for stronger unity between the former Spanish colonies of South America
 C. Slave rebellions fragmenting newly independent Latin American states
 D. Sectionalism of elites in the various territories of newly independent Latin American states
Ground Truth: D

2005
 2006
 2007
 2008
 2009
 2010
 2011
 2012
 2013
 2014
 2015
 2016
 2017
 2018
Reasoning:
 To determine which option most undermines the expectations expressed in Article 1 of Source 1, let's analyze the article and the options step by step.
 Article 1 of Source 1 states:
"The Republics of Venezuela and New Granada are from this day on united in a single one under the glorious title of Republic of Colombia."
 This article expresses the expectation that the Republics of Venezuela and New Granada should be united to elevate themselves to a higher grade of power and prosperity.
Let's search for a fact to help us determine which option most undermines this expectation.
 <se>
What was the outcome of the union between Venezuela and New Granada?
 </se>
 <info>
 The union between Venezuela and New Granada, along with Ecuador and Panama, resulted in the creation of the Republic of Gran Colombia in 1819....
 ...
Based on the information provided, the correct option that most undermines the expectations expressed in Article 1 of Source 1 is the one that relates to the fragmentation of the union.
Option A:
"Intervention by the United States and European powers in the independence of Latin American countries"
 This might have influenced the political landscape but does not specifically address the fragmentation of the union.
 ...
This is more likely to be a factor in the fragmentation of the union, but let's check the fact again.
Let's refine:
 <se>
What internal conflicts led to the dissolution of Gran Colombia?
 </se>
 <info>
 The dissolution of Gran Colombia in 1830–1831 was primarily caused by internal political conflicts and regional tensions. The main internal conflicts included: ...
 In summary, the dissolution of Gran Colombia was driven by internal political divisions over federalism versus centralism, strong regional identities and rivalries, and leadership struggles, all of which led to civil conflict and the eventual fragmentation of the union.
 </info>
 Based on the information provided, option D seems to be the best fit.
Thus, the correct answer is: D

2044
 2045
 2046
 2047
 2048
 2049
 2050
 2051

Refined Query

Figure 9: UR² initially retrieved general information about Gran Colombia's dissolution, but upon reflection recognized this was insufficient for precise answer selection. It then refined the query to investigate specific internal conflicts, ultimately identifying elite sectionalism as the key undermining factor, demonstrating effective use of Reflection through query refinement.

2052
2053
2054
2055
2056
2057
2058

Reflection (Retry Retrieve/Refine Query)

Question:

Nathanson supports which of the following kinds of retributivism?

Options:

- A. proportional
- B. equality
- C. retaliatory
- D. punitive
- E. compensatory
- F. absolute
- G. restorative
- H. corrective
- I. preventive
- J. consequential

Ground Truth: A

Reasoning:

To determine which kind of retributivism Nathanson supports, let's analyze the concept of retributivism and the options provided.

First, let's search for specific information about Nathanson's stance on retributivism.

<se>

What kind of retributivism does Nathanson support?

</se>

<info>

The provided Wikipedia content does not mention Nathanson or specify what kind of retributivism he supports. However, it does define retributive justice as a theory of punishment that holds offenders should suffer in proportion to their offense, emphasizing proportionality and procedural standards, and distinguishing retribution from revenge. Without direct reference to Nathanson, no specific information about his stance or the particular kind of retributivism he supports can be extracted from the current Wikipedia content.

</info>

Based on the information provided, it seems we need more specific details about Nathanson's stance. Let's try another approach by looking at the general types of retributivism and see if any of them fit the description of Nathanson's support.

<se>

What are the main types of retributivism in criminal justice?

</se>

<info>

The Wikipedia content indicates that retributivism is a theory of punishment in criminal justice which holds that offenders deserve to suffer in proportion to their offenses. While the content does not list explicit "types" of retributivism by name, it does reference several key distinctions and purposes within retributive theories:

- **Proportionality**: Retributive justice emphasizes that the punishment should be proportional to the crime committed.
- **Desert-Based (Backward-Looking) Retributivism**: This form holds that offenders deserve punishment simply because they have committed a wrongdoing, focusing on blameworthiness and restoring balance (e.g., Lex Talionis or the Unfair Advantage Principle).
- **Expressive or Denunciatory Retributivism**: This type emphasizes the role of punishment in openly denouncing crime and expressing societal condemnation.

- **Restorative Elements**: Some retributive theories aim to restore balance or provide satisfaction to victims or society.

Additionally, the content distinguishes retributive justice from revenge by noting that retribution is not personal, has inherent limits, and follows procedural standards.

In summary, the main types or approaches within retributivism in criminal justice, as reflected in the Wikipedia content, include:

- Proportional (or "just deserts") retributivism
- Expressive/denunciatory retributivism
- Restorative or balance-restoring retributivism

</info>

Based on the information provided, let's compare the options with the types of retributivism mentioned:

Option A, proportional, seems to be the best fit based on the detailed description provided.

Thus, the correct answer is: A

Retry & Refine Query

2098
2099
2100
2101
2102
2103
2104
2105

Figure 10: This case demonstrates UR² capacity to reflect and refine its query strategy when initial retrieval fails. By shifting from a direct search for Nathanson's view to analyzing types of retributivism, the model successfully deduced the answer through indirect reasoning. This exemplifies its performance in reflection and retry-based retrieval refinement.