ADATUTORANK: LEARNING TO RERANK DOCUMENT SETS VIA ADAPTIVE TUTORING OPTIMIZATION FOR RAG AND DEEP RESEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Document rerankers determine what evidence reaches the downstream model in RAG and deep research, yet mainstream rerankers select by relevance matching, and individually relevant documents rarely constitute the complete, complementary, non-redundant set a complex information need demands. Prior work rewards a set by its aggregate rubric score, shifting the objective from ranking documents to composing sets. Yet that score is one scalar shared by every document in the set, so the supervision is sparse: a redundant document is rewarded with the rest whenever the set scores well, and a decisive one penalized with the rest whenever it does not; credit assignment leaves *contributors* indistinguishable from *free riders*. On-policy distillation could densify this supervision, but existing methods give every rollout the same fixed guidance, too prescriptive for strong rollouts and too abstract for weak ones. We therefore propose **AdaTutoRank**, a setwise reranker trained with **Ada**ptive **Tuto**ring Optimization (ATO) under a three-level hierarchy of nine rubric dimensions, which supplies silver labels for the cold start, rewards for reinforcement learning, and hints for distillation. ATO draws three hint forms of increasing specificity from the policy's own frozen snapshot: the rubrics alone, a *self-selector*'s sibling-set chosen under rubrics, and a *self-reflector*'s reflection contrasting the rollout with that sibling-set; each rollout receives the form matched to its quality. Re-scoring that rollout under the hint-conditioned frozen teacher and the hint-free snapshot distills the hint's effect into a token-level advantage that complements the group-relative outcome advantage. Across ten benchmarks spanning RAG, deep research, and setwise evaluation, AdaTutoRank attains the best overall performance while issuing fewer retrieval calls.

1 INTRODUCTION

Document rerankers determine the quality of the evidence passed to the downstream LLM in RAG and deep research, and thus the accuracy of the generated answer (Gao et al., 2023; Liu et al., 2023; Yoran et al., 2024; Li et al., 2025; Xu & Peng, 2025; Shi et al., 2025). A deep research agent further interacts with the search system over multiple steps, reasoning about its current information need, issuing a sub-query, observing the returned documents, and deciding whether to keep searching before producing a long-form answer (Yao et al., 2022; Jin et al., 2025; Shao et al., 2025). Retrieval quality therefore governs not only the current observation but also the reasoning and search decisions that follow, so a single poor observation propagates and compounds over the trajectory.

Mainstream rerankers still ground their supervision in relevance annotations and return the top-k individually relevant documents (Zhuang et al., 2022; Xiao et al., 2023). Yet the downstream model needs not a collection of such documents, but a set that jointly supports answering the query, covering it comprehensively without redundancy or conflict. As shown in Figure 1, for the question "Explain the pros and cons of using Blockchain for supply chain visibility", a relevance-based reranker returns documents that are all on topic yet form a poor set. ❶ They merely match the topic keywords rather than answer the question. ❷ Doc_1 and Doc_2 restate the same point, consuming context budget without adding information. ❸ All three speak only to the benefits, leaving the limitations uncovered.

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

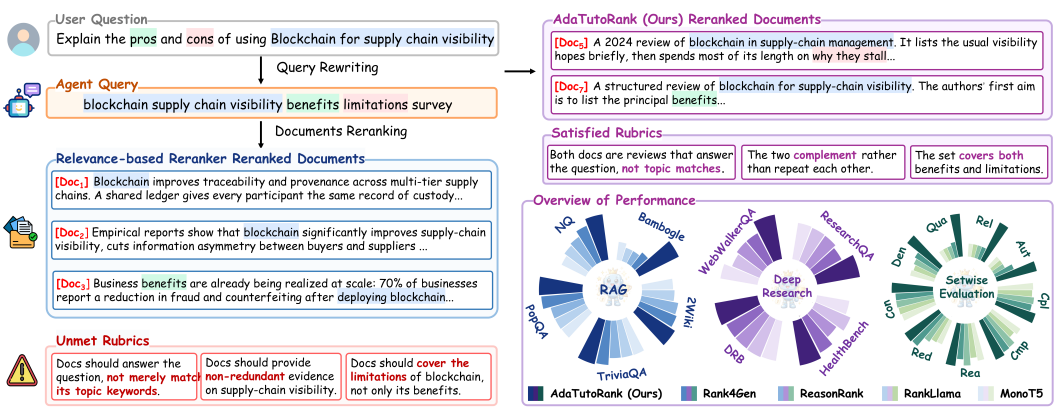


Figure 1: **(i)** Relevance-based reranking returns individually on-topic documents whose set is superficially relevant, redundant, and incomplete. **(ii)** AdaTutoRank instead selects a set that is genuinely relevant, complementary, and comprehensive, while consistently performing best on RAG, deep research, and setwise evaluation.

Recent work has shown the promise of rubrics for guiding document-set selection. RubricRanker (Liu et al., 2026a) drives GRPO training with a rubric-based reward, while Rubric4Setwise (Jiang et al., 2026) employs rubrics as a training-free prompting signal. Instantiating a rubric as an RL reward, however, collapses the supervision into a set-level scalar spread uniformly over every document. Three problems follow. ❶ Document-level credit assignment is absent, as redundant documents free-riding on a high-reward set go unpunished, while decisive documents in a low-reward set are penalized with the rest. ❷ Without dense process supervision, the scalar invites reward hacking: selecting a single document trivially incurs zero redundancy and zero conflict, scoring highly on both. ❸ Its rubrics cover mainly relevance, conflict, and redundancy, leaving density, completeness, and complementarity unmeasured, which compounds the two problems above.

The first two problems both stem from the sparsity of the reward. On-policy distillation is well suited to remedying it, since rescoring the same rollout under a context augmented with privileged information yields dense supervision resolved to individual documents. Prior work has instantiated such information as environment feedback, reference solutions, demonstration trajectories, or transient experience, yet all share one limitation (Hubotter et al., 2026; Zhao et al., 2026; Shenfeld et al., 2026; Ye et al., 2026). Although its content varies across samples, its type is fixed for the entire training set, which teaches every sample with one teacher and one recipe and cannot adapt to rollouts that fail in different ways. Effective supervision over document sets should therefore satisfy three requirements.

First, the supervision should be ***on-policy***, because useful corrections depend on the failure modes the current policy exhibits, whereas fixed demonstrations, static reference sets, and one-time distillation cannot track a policy whose capability and outputs evolve. Second, it should be ***dense***, because a scalar reward says only whether the whole set is good, not which document to keep and which to drop. Third, it should be ***adaptive tutoring***, because rollouts differ in why they fail and how much correction they can absorb, so the privileged information should vary in form with rollout quality.

We propose **AdaTutoRank**, a setwise reranker trained under rubric supervision with *Adaptive Tutoring Optimization* (ATO). We introduce rubrics organized as a three-level hierarchy of nine dimensions and carry them through the entire training process, supplying silver labels for supervised fine-tuning and both the scalar reward and the privileged information for ATO. Training proceeds in two stages. *Setwise Supervised Fine-Tuning* fine-tunes the policy on those silver labels, giving a stable cold start that selects sets in the required form. *Adaptive Tutoring Optimization* then supplies the dense signal that the scalar reward lacks. At each step the frozen policy generates a hint for every rollout, matched to its reward. ❶ A high-reward rollout receives the rubrics alone, encouraging closer conformance; ❷ a medium-reward one receives a reflection contrasting its own set with the rubric-selected set, correcting erroneous selections; and ❸ a low-reward one receives that sibling-set directly as a correction. Re-scoring a rollout under its hint turns the hint's effect into a token-level advantage, which is combined with the group-relative outcome advantage so that the quality of the set and that of each document within it are optimized together. Hints and query-specific rubrics are used only in training, and at inference the policy sees only the query, meta-rubric and candidates.

The contributions of this paper are summarized as follows:

- We introduce rubrics organized as a three-level hierarchy of nine dimensions and carry them through the training pipeline, where they supply silver labels, reinforcement rewards, and distillation hints, making multi-dimensional set quality explicitly optimizable at every stage.
- We propose adaptive tutoring optimization, which matches the form of the hint to each rollout's quality and distills its effect into a token-level advantage combined with the group-relative outcome advantage, turning set-level utility into token-level credit.
- Across ten benchmarks spanning answer-level and setwise-level evaluation, AdaTutoRank achieves the best overall performance while reducing the agent's retrieval calls.

2 Related Work

RAG and Deep Research. Retrieval-augmented generation (RAG) lets LLMs draw on knowledge beyond their parameters (Gao et al., 2023; Deng et al., 2026), yet a single retrieval round rarely suffices for complex information-seeking tasks. This has motivated agentic search (Yao et al., 2022) and deep research agents that interleave reasoning with web search for open-ended queries (Shi et al., 2025; Li et al., 2025). Such agents change what a reranker serves: it is invoked once per self-issued sub-query rather than per user question, and its output becomes the observation conditioning the next reasoning step, so any defect in that set propagates through every step that follows.

Document Reranking. Rerankers have evolved from ordering documents toward composing sets. Ad hoc methods rank candidates by relevance with trained scorers (Xiao et al., 2023; Nogueira et al., 2020; Ma et al., 2023) or LLM prompting (Pradeep et al., 2023a;b), and reasoning-enhanced methods add explicit reasoning (Zhang et al., 2025; Liu et al., 2026b). Both optimize pointwise ordering, leaving set composition to a fixed cutoff. Setwise methods instead predict the subset directly, supervised by answer-generation quality (Fan et al., 2026) or a strong teacher (Lee et al., 2025), yet neither exists for open-ended queries with unverifiable answers. Rubrics fill this gap with structured criteria, yet RubricRanker (Liu et al., 2026a) trains on a set's aggregate rubric score, a single set-level scalar from which the gradient cannot separate contributors from free riders.

Prior rubric-based training shares one scalar across the whole set, crediting no document individually. AdaTutoRank turns the rubrics into hints, matches their form to rollout quality, and distills each into token-level shaping over the identifier sequence.

3 Methodology

We propose **AdaTutoRank**, a setwise reranker for RAG and deep research, trained under rubric supervision throughout the pipeline with **Ada**ptive **Tuto**ring Optimization (ATO). Its design follows two observations: a rubric-based reward is a single scalar spread uniformly over the set and thus carries no document-level credit; and one fixed type of privileged information cannot serve rollouts that fail for different reasons. ATO therefore supplements the sparse reward with adaptive tutoring hints, dense supervision whose form varies with rollout quality.

As shown in Figure 2, this yields two training stages. We first synthesize query-specific rubrics, use them to produce silver labels, and fine-tune the policy into a stable cold start. We then distill the hints into token-level supervision and combine it with group-relative outcome advantages.

3.1 Preliminary

Given a query q and a candidate pool $\mathcal{C} = \{d_1, \ldots, d_m\}$ returned by a retriever, a reranker decides which documents serve as evidence for answering q. Most rerankers learn a scoring function $\sigma(q, d)$ measuring the relevance of d to q, and truncate the ranking it induces over $\mathcal{C}$ at a fixed cutoff k, *i.e.,* $\mathcal{S} = \mathrm{Top}_k\big(\sigma(q, \cdot)\big)$, so a hyperparameter rather than the model determines what the evidence contains. We therefore adopt the setwise formulation, in which the subset itself is the prediction,

$$\mathcal{S}^\star = \arg\max_{\mathcal{S} \subseteq \mathcal{C}} u(\mathcal{S} \mid q), \tag{1}$$

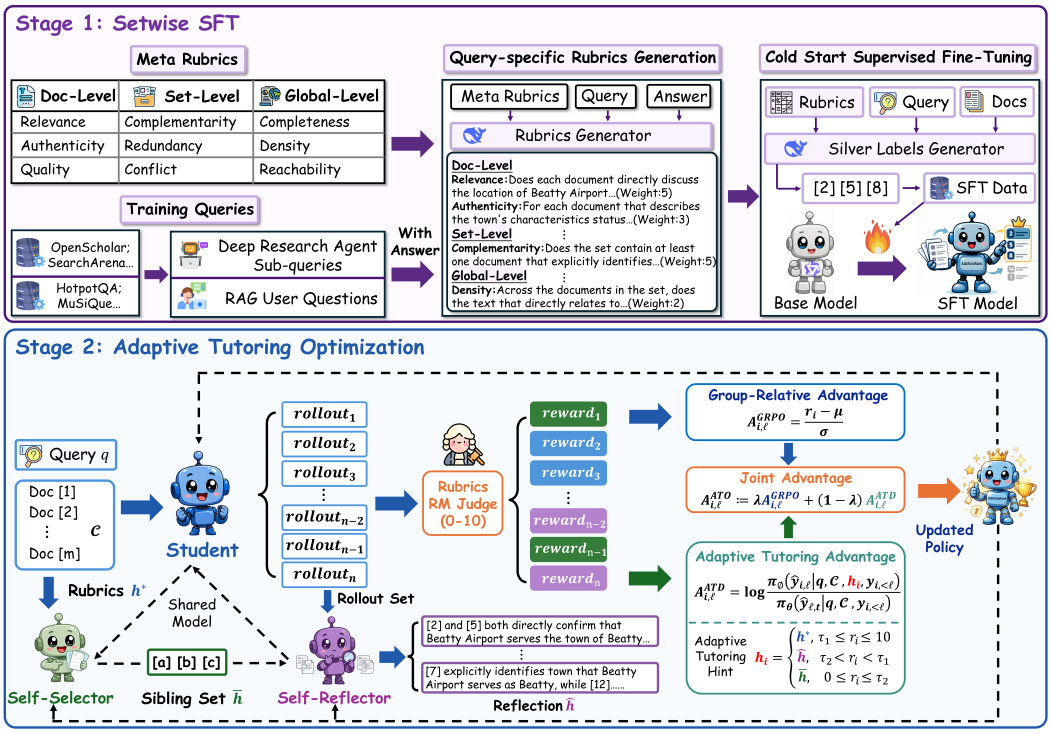


Figure 2: **Overview of AdaTutoRank. Stage 1** (Setwise SFT) generates silver labels from query-specific rubrics and fine-tunes the policy into a cold start for set selection. **Stage 2** (Adaptive Tutoring Optimization) routes to each rollout a hint matched to its reward, and re-scoring the rollout with and without that hint yields a dense advantage complementing the group-relative outcome signal.

where the set utility $u(\cdot \mid q)$ is non-additive over documents and $|\mathcal{S}^\star|$ is set by the information need of q rather than by a hyperparameter. The reranker thereby takes on two decisions that ranking leaves out, ❶ *which documents belong together* and ❷ *how many are enough*.

Since u is not directly observable, we instantiate it with the query-specific rubrics: a reward model scores $\mathcal{S}$ against each rubric in R_q and aggregates them into the training-time value of $u(\mathcal{S} \mid q)$ (Eq. 5). The reranker is an autoregressive policy that outputs the selected identifiers directly,

$$y \sim \pi_\theta(\cdot \mid q, \mathcal{C}), \qquad \mathcal{S} = \mathrm{Parse}(y), \tag{2}$$

where y is a bracketed identifier sequence such as `"[2] [5] [8]"`.

The formulation covers both scenarios, which differ only in what q is and how often the policy runs. In RAG, it runs once on the user question, so $\mathcal{S}$ is the generator's entire evidence and $u(\mathcal{S} \mid q)$ bounds answer quality. In deep research, the agent (Shao et al., 2025) follows ReAct (Yao et al., 2022) and interleaves reasoning with search; the policy runs after every search action, with $q = q_t$ a self-issued sub-query and $\mathcal{S}_t \subseteq \mathcal{C}_t$ becoming that step's observation.

3.2 Rubrics Construction

Following prior works (Liu et al., 2026a; Jiang et al., 2026), we construct query-specific rubrics specifying the properties a selected document set should satisfy at the doc, set, and global levels.

Meta Rubrics Design. Prompting an LLM for rubrics directly tends to yield limited coverage, repetitive items, or conflated dimensions. RubricRanker mitigates this with a fixed schema, but its five dimensions are too narrow to characterize a document set. We therefore adopt meta rubrics organized as a three-level hierarchy of nine dimensions. The document level focuses on each document in isolation, through ❶ *Relevance* (Rel.), ❷ *Authenticity* (Aut.), and ❸ *Quality* (Qua.); the set level focuses on the synergy between documents, through ❹ *Complementarity* (Cmp.), ❺ *Redundancy* (Red.), and ❻ *Conflict* (Con.); and the global level focuses on the set as LLM input context, through ❼ *Completeness* (Cpl.), ❽ *Density* (Den.), and ❾ *Reachability* (Rea.), with details in Appendix C.1.

Training Query Collection. For RAG, we take short closed-ended questions from HotpotQA (Yang et al., 2018), NQ (Kwiatkowski et al., 2019), 2WikiMultihopQA (Ho et al., 2020), and MuSiQue (Trivedi et al., 2022). In deep research, a reranker serves agent-issued sub-queries, not user questions, which open-source datasets rarely release. We therefore reuse those from RubricRanker, covering OpenScholar (Asai et al., 2024), SearchArena (Miroyan et al., 2026), GlaiveAI-Reasoning-v1-20M, and WebWalker-Silver (Wu et al., 2025), with details in Appendix C.2.

Query-Specific Rubrics Generation. We instantiate R_q by prompting DeepSeek-V4 Pro with the meta rubrics, the query q, and its reference answer a, producing one or more rubrics per dimension. Each rubric is an evaluation question on set quality that must cite specific entities, facts, or values from q and a; vague phrasing such as "relevant content" is prohibited.

3.3 Setwise Supervised Fine-Tuning

The first stage cold-starts the policy to select document sets that jointly satisfy the multi-dimensional criteria in R_q, supervised by silver labels that a frontier LLM produces conditioned on q and R_q.

Silver Label Generation. Since a base model rarely assembles a high-quality set on its own, its rollouts would impede the subsequent optimization stage; we thus cold-start the policy on silver labels from DeepSeek-V4-Pro. Conditioned on q, $\mathcal{C}$, and R_q, it emits y^*, the identifiers of the retained documents, with $|\mathcal{C}|$ sampled from 10 to 40 per query to expose it to varying pool sizes.

Setwise SFT. We then fine-tune the policy to predict y^* from q and $\mathcal{C}$ alone. R_q is withheld, since it depends on a reference answer unavailable at inference, so training and inference see identical inputs. Each example $(q, \mathcal{C}, y^*) \in \mathcal{D}_{\text{sft}}$ is optimized with the negative log-likelihood,

$$\mathcal{L}_{\text{sft}}(\theta) = -\mathbb{E}_{(q,\mathcal{C},y^*)\sim\mathcal{D}_{\text{sft}}}\left[\sum_{\ell=1}^{|y^*|} \log \pi_\theta(y^*_\ell \mid q, \mathcal{C}, y^*_{<\ell})\right], \tag{3}$$

where ℓ indexes the tokens of y^*. The resulting θ_{sft} initializes the next-stage policy and serves as its frozen distillation teacher (Eq. 9).

3.4 Adaptive Tutoring Optimization

The second stage performs adaptive tutoring optimization. At each update, we freeze the current policy as $\pi_{\theta_{\text{old}}}$, which both samples the rollout group $\{y_i\}_{i=1}^{G}$ and parameterizes the two hint generators of Figure 2, a *self-selector* and a *self-reflector*. Each rollout is then paired with the hint matched to its reward r_i, and π_θ is optimized jointly with the reinforcement learning and distillation objectives before becoming the next snapshot. Supervision therefore stays on-policy: hints derive from the policy's own rollouts, and their effect is distilled back into a policy that needs no hint at inference.

Hierarchical Rubric-based Reward. Given the query-specific rubrics $R_q = \{R_1^{\text{doc}}, \ldots, R_x^{\text{doc}}\} \cup \{R_1^{\text{set}}, \ldots, R_y^{\text{set}}\} \cup \{R_1^{\text{global}}, \ldots, R_z^{\text{global}}\}$, where R_i^{doc}, R_j^{set}, and R_k^{global} denote the i-th doc-level, j-th set-level, and k-th global-level rubric with weights w_i^{doc}, w_j^{set}, and w_k^{global} assigned during rubric construction, we compute the reward through a hierarchical aggregation with these weights.

The reward is provided by a rubric-based judge J, instantiated with DeepSeek-V4-Flash, which scores the selected set $\mathcal{S}$ against each rubric in R_q on a 0–10 scale. Set and global-level rubrics are rated once for the whole set, as $J(\cdot, \mathcal{S})$, whereas doc-level rubrics are rated per doc and averaged,

$$F(R_i^{\text{doc}}, \mathcal{S}) = \frac{1}{|\mathcal{S}|}\sum\nolimits_{d\in\mathcal{S}} J(R_i^{\text{doc}}, d). \tag{4}$$

The three levels are then aggregated by their weights, instantiating the set utility of Eq. 1,

$$u(\mathcal{S} \mid q) = \frac{\sum_i w_i^{\text{doc}} \cdot F(R_i^{\text{doc}}, \mathcal{S}) + \sum_j w_j^{\text{set}} \cdot J(R_j^{\text{set}}, \mathcal{S}) + \sum_k w_k^{\text{global}} \cdot J(R_k^{\text{global}}, \mathcal{S})}{\sum_i w_i^{\text{doc}} + \sum_j w_j^{\text{set}} + \sum_k w_k^{\text{global}}}, \tag{5}$$

We further require the output to be a bracketed identifier sequence, e.g., `[2] [5] [8]`. The reward of rollout y_i is then the utility of its parsed set $\mathcal{S}_i = \text{Parse}(y_i)$ if the format is valid, and -1 otherwise,

$$r_i = \begin{cases} u(\mathcal{S}_i \mid q) \in [0, 10], & \text{if the output format is correct,} \\ -1, & \text{otherwise.} \end{cases} \tag{6}$$

Following GRPO (Shao et al., 2024), we then obtain the reinforcement advantage by standardizing r_i within the group of G rollouts sharing the same query,

$$A_{i,\ell}^{\mathrm{GRPO}} := \frac{r_i - \mathrm{mean}\{r_{i'}\}_{i'=1}^{G}}{\mathrm{std}\{r_{i'}\}_{i'=1}^{G}} \quad (\text{constant in } \ell), \tag{7}$$

where ℓ indexes tokens of y_i. This advantage reflects the outcome but is uniform across the rollout, providing no token-level supervision. We thus introduce a distillation signal adapted to rollout quality.

Adaptive Tutoring Hints. A hint is privileged information used only in the distillation branch, where re-scoring y_i under it yields the missing token-level signal. Prior methods use one hint type throughout training, so identical guidance is ❶ *too prescriptive for strong rollouts* or ❷ *too abstract for poor ones*. We therefore match hint form to rollout quality.

At the beginning of each training step, the frozen snapshot $\pi_{\theta_{\mathrm{old}}}$ drives two hint generators. The *self-selector* maps $(q, \mathcal{C}, R_q)$ to a sibling-set $\bar{h}$, an alternative document set chosen from $\mathcal{C}$ under the rubrics; the *self-reflector* maps $(q, \mathcal{C}, y_i, \bar{h})$ to a reflection $\hat{h}$ diagnosing which documents y_i wrongly retained or missed relative to $\bar{h}$. With the rubric hint $h^* = R_q$, which states the criteria alone, these three forms constitute the hint space, and Figure 3 shows how each is injected into the teacher prompt. Which form a rollout receives follows from its reward r_i through two thresholds $\tau_1 > \tau_2$,

$$\boldsymbol{h}_i = \begin{cases} h^*, & \tau_1 \le r_i \le 10, \\ \hat{h}, & \tau_2 < r_i < \tau_1, \\ \bar{h}, & 0 \le r_i \le \tau_2. \end{cases} \tag{8}$$

The tiers differ in how much correction a rollout can absorb. A high-reward rollout is already near a qualifying set, so h^* refines it without overriding its justified choices; a mid-reward one is partly right, so $\hat{h}$ corrects only its errors; a low-reward one offers little to build on and is replaced outright by $\bar{h}$.

Sibling-set Gate. Both $\hat{h}$ and $\bar{h}$ build on the sibling-set, so we gate it with the same judge: $\bar{h}$ is delivered only when it outscores the rollout it supervises, $u(\bar{h} \mid q) > r_i$; otherwise the rollout falls back to h^*, and no hint ever teaches toward a worse reference. We set $\tau_1 = 6$ and $\tau_2 = 3$ throughout.

Joint Training Objective. Beyond the outcome-driven GRPO advantage, we supervise the policy with a token-level *adaptive tutoring distillation* (ATD) advantage derived from the hint. Without regenerating y_i, re-scoring it under $(q, \mathcal{C}, \boldsymbol{h}_i)$ and $(q, \mathcal{C})$ yields, for its ℓ-th token $\hat{y}_{i,\ell}$,

$$A_{i,\ell}^{\mathrm{ATD}} := \log \frac{\pi_{\phi}\big(\hat{y}_{i,\ell} \mid q, \mathcal{C}, \boldsymbol{h}_i, y_{i,<\ell}\big)}{\pi_{\theta_{\mathrm{old}}}\big(\hat{y}_{i,\ell} \mid q, \mathcal{C}, y_{i,<\ell}\big)}, \tag{9}$$

where the hint-conditioned teacher π_ϕ is the cold-start checkpoint θ_{sft}, frozen throughout the stage. Were it to move with the policy, the distillation branch would lose its external reference and risk reinforcing the policy's own mistakes. Since ϕ and θ_{old} are both fixed within the update, $A_{i,\ell}^{\mathrm{ATD}}$ is constant in θ and enters purely as an advantage, positive where the hint makes a token more likely under the teacher; its sign anchors the policy to θ_{sft}, obviating an explicit KL penalty.

The final ATO advantage combines group-relative outcome feedback with token-level supervision,

$$A_{i,\ell}^{\mathrm{ATO}} := \lambda A_{i,\ell}^{\mathrm{GRPO}} + (1-\lambda) A_{i,\ell}^{\mathrm{ATD}}, \qquad \lambda \in [0,1], \tag{10}$$

where λ balances the two signals and is set to 0.7 in our experiment.

This formulation keeps the outcome reward as the primary RL signal while adding token-level shaping. We optimize the standard clipped policy objective on the combined advantage,

$$\mathcal{L}_{\mathrm{policy}}(\theta) = -\mathbb{E}_{i,\ell}\Big[\min\big(\rho_{i,\ell}(\theta)\, A_{i,\ell}^{\mathrm{ATO}},\ \mathrm{clip}(\rho_{i,\ell}(\theta), 1-\epsilon, 1+\epsilon)\, A_{i,\ell}^{\mathrm{ATO}}\big)\Big], \tag{11}$$

where $\rho_{i,\ell}(\theta)$ denotes the token-level importance ratio, defined as

$$\rho_{i,\ell}(\theta) = \exp\big(\log \pi_\theta(\hat{y}_{i,\ell} \mid q, \mathcal{C}, y_{i,<\ell}) - \log \pi_{\theta_{\mathrm{old}}}(\hat{y}_{i,\ell} \mid q, \mathcal{C}, y_{i,<\ell})\big). \tag{12}$$

Here clip bounds the importance ratio to $[1-\epsilon, 1+\epsilon]$, with ϵ limiting the deviation from the old policy. As noted above, the anchoring in Eq. 9 lets us omit the KL penalty.

Training-inference boundary. The self-selector, self-reflector, and hint serve only to construct the training advantage. At inference the policy acts from $(q, \mathcal{C})$ and the fixed meta-rubric instruction, plus the query intent for deep research, without the query-specific R_q or its reference answer.

Student Prompt

System prompt. You are a high-quality document set selector for a deep-research agent. {task_description}; {meta_rubrics}; {output_format}.

User prompt. I will provide you with {num_documents} documents, each indicated by a numerical identifier []. Select and rank the documents based on their joint usefulness for the search query: {query}; {query_intent}; {candidate_documents}.

Documents selection:

Teacher Prompt

System prompt & User prompt. Identical to the student prompt above.

Adaptive Tutoring Hints. One of the three, adaptively matched to the reward of the rollout.

[Rubrics as hint] Below are the scoring dimensions and key points that the optimal document subset for this query should cover (a higher weight means more important): {rubric}

```
[Doc-level rubrics]
- Relevance (weight 5): Does each document give the main role and archetype of the specific champions
listed in the query, rather than discussing general class systems or champion lore?

⋮ more rubrics across the other dimensions
```

Strictly follow the rubrics above and select the optimal document subset.

[Corrective reflection as hint] Your current selection is not yet optimal. Below is a concrete corrective idea for revising it toward the optimal document subset: {corrective_reflection}

```
[2] and [22] should be included: [2] gives the structured class system and [22] the specific subclass
assignments. [1] [5] [7] [19] [32] are less effective, as they contain unrelated content, repeat one
another, or do not directly cover many champions.
```

Apply the corrective idea above and select the optimal document subset.

[Sibling-set as hint] Below is an example optimal document subset for this query (scored {reward} out of 10 by the judge); use it as a concrete reference: {sibling_set}

```
[2] [22]
```

Learn from the reference example above and select the optimal document subset.

Documents selection:

Figure 3: Prompt example for the student and teacher policies.

4 EXPERIMENTS

4.1 SETTINGS

Benchmarks. We evaluate AdaTutoRank at two granularities. ❶ **Answer-level evaluation** assesses the final response under two scenarios. For **RAG**, we use Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), PopQA (Mallen et al., 2023), 2WikiMultihopQA (Ho et al., 2020), and Bamboogle (Press et al., 2023), whose short closed-ended answers are measured by exact match (EM). For **deep research**, we use WebWalkerQA (Wu et al., 2025), HealthBench (Arora et al., 2025), DeepResearchBench (Du et al., 2026), and ResearchQA (Yifei et al., 2026), whose open-ended responses are scored by LLM judges. ❷ **Setwise-level evaluation** assesses the document set itself with SetwiseEvalKit (Jiang et al., 2026), a rubric-based benchmark scoring a set along nine dimensions beyond relevance (*e.g.,* completeness, redundancy, conflict).

Baselines. We compare AdaTutoRank with 11 rerankers in three categories. ❶ **Adhoc reranking**, which orders documents by pointwise relevance: BGE-Reranker-Large (Xiao et al., 2023), MonoT5 (Nogueira et al., 2020), RankT5 (Zhuang et al., 2022), RankLlama (Ma et al., 2023), RankVicuna (Pradeep et al., 2023a), and RankZephyr (Pradeep et al., 2023b). ❷ **Reasoning-enhanced reranking**, which adds explicit reasoning to the decision: Rearank (Zhang et al., 2025) and ReasonRank (Liu et al., 2026b). ❸ **Setwise reranking**, which selects a set rather than a ranked list: Rank4Gen (Fan et al., 2026), SetR (Lee et al., 2025), and RubricRanker (Liu et al., 2026a). We also report Initial Retrieval, the unreranked top-5 from BM25 (Robertson & Zaragoza, 2009) or the Google Search API, as a lower bound.

Table 1: **Answer-Level Performance Comparison.** All rerankers rerank the top-20 documents, and we report exact match and LLM-judge scores on RAG and deep research respectively. The **best** and second-best results are highlighted. All metrics are reported such that higher is better (↑).

Ranker	RAG						Deep Research					Overall
	Avg.	NQ	Triv	PopQA	2Wiki	Bambo	Avg.	WebW	HealthB	DRB	RQA	
Only Retrieval (BM25 & Google Search API)												
Initial Retrieval	31.95	28.45	60.05	32.72	27.35	11.20	47.98	32.00	45.40	48.62	65.89	39.08
Adhoc Reranking												
BGE-Reranker$_{L.}$	35.16	32.94	62.80	35.85	29.82	14.40	51.01	**42.00**	45.60	48.84	67.59	42.20
MonoT5 (3B)	34.42	31.25	62.38	35.19	28.87	14.40	49.73	39.00	44.09	48.01	67.81	41.22
RankT5 (3B)	34.13	31.75	62.34	35.47	29.07	12.00	49.70	40.00	43.57	48.15	67.07	41.05
RankLlama (7B)	33.77	32.38	62.17	35.40	29.32	9.60	49.55	36.00	46.44	48.35	67.39	40.78
RankVicuna (7B)	34.61	32.58	61.85	35.10	28.30	15.20	50.34	35.00	46.92	48.39	71.04	41.60
RankZephyr (7B)	34.89	33.68	62.19	35.32	28.08	15.20	50.55	41.00	44.80	47.91	68.48	41.85
Reasoning-Enhanced Reranking												
Rearank (7B)	35.11	32.41	62.31	35.31	30.32	15.20	49.16	36.00	44.70	47.64	68.29	41.35
ReasonRank (7B)	34.98	33.02	62.09	35.49	30.72	13.60	48.83	37.00	43.00	48.12	67.18	41.14
Setwise Reranking												
Rank4Gen (8B)	37.00	34.93	65.76	37.87	31.23	15.20	49.26	37.00	46.60	48.15	65.30	42.45
SetR (8B)	36.87	34.35	65.64	37.32	31.85	15.20	50.96	40.00	46.00	49.31	68.54	43.13
RubricRanker (8B)	38.25	35.68	66.30	37.65	31.61	20.00	50.83	37.00	45.90	**49.95**	70.48	43.84
AdaTutoRank (8B)	**39.05**	**36.87**	**66.54**	**38.52**	**32.52**	**20.80**	**53.06**	41.00	**48.57**	49.89	**72.78**	**45.28**

Implementation Details. AdaTutoRank uses Qwen3-8B (Yang et al., 2025) as its backbone. Documents come from the Google Search API on deep research benchmarks and from BM25 over the Wikipedia dump (Karpukhin et al., 2020) on RAG benchmarks. Every reranker operates on the top-20 retrieved documents and returns a set. More details are provided in Appendix B and C.

4.2 Analysis of Experimental Results

Analysis of Main Results. Tables 1 and 2 report AdaTutoRank against baselines at two granularities. Answer-level results show whether the downstream task improves, but confound the generator with the evidence it receives; setwise-level results isolate the evidence, revealing whether gains originate from a better selected set. We draw three observations: ❶ ***AdaTutoRank improves downstream answer performance across the board.*** In Table 1, it attains an overall score of 45.28, exceeding the strongest baseline RubricRanker by 1.44 ↑ points and Initial Retrieval by 6.20 ↑, and it ranks first on seven of the nine benchmarks and second on the remaining two. ❷ ***AdaTutoRank yields larger gains in deep research than in RAG.*** In Table 1, it raises the scenario average by 0.80 ↑ on RAG (39.05 vs. 38.25) and by 2.05 ↑ on deep research (53.06 vs. 51.01). This follows from how far the evidence acts: in RAG the reranker runs once, its effect confined to a single generation step, whereas in deep research each set becomes the observation for the next sub-query, so better evidence compounds over the trajectory. ❸ ***AdaTutoRank attains superior evidence quality on set selection.*** In Table 2, it attains the highest average at all three rubric levels, ranks first on seven of the nine fine-grained dimensions, and lifts the overall score by 2.17 points over the strongest baseline RubricRanker. As these sets are what the generator and the agent consume, the gains in Table 1 trace to the evidence rather than its consumer, resolving the confound above.

Analysis of Ablation Experiment Results. We ablate four aspects of AdaTutoRank, with results reported in Table 3, and summarize our observations as follows. ❶ ***Training Stage.*** We ablate the two-stage framework by removing ATO ("w/o ATO") and cold-start SFT ("w/o SFT"). Dropping ATO lowers the overall score from 49.19 to 46.33, while dropping SFT lowers it to 47.04. Removing either stage costs a comparable amount, indicating that the cold start and ATO are complementary rather than substitutable. ❷ ***Advantage Superposition.*** Optimizing with the outcome advantage alone ("Only RL") yields 45.07, even below the 46.33 of SFT alone, indicating that a scalar reward shared by every selected identifier is too coarse to improve set selection. The distillation advantage alone ("Only Distillation") reaches 47.79, and superposing the two as in Eq. 10 gives the best 49.19, confirming that the two signals are complementary. ❸ ***Interpolation Coefficient.*** Performance degrades on both sides of $\lambda = 0.7$ (49.19), falling to 47.75 at $\lambda = 0.5$ and 46.61 at $\lambda = 0.9$. Each end collapses toward the corresponding single-signal regime: a large λ suppresses the tutoring advantage and drifts toward "Only RL" (45.07), while a small one weakens the outcome grounding

Table 2: **Setwise-Level Performance Comparison on SetwiseEvalKit (Short-form Scenario).** All metrics are reported such that higher is better (↑).

Ranker	Doc-Level				Set-Level				Global-Level				Overall
	Avg	Rel.	Aut.	Qua.	Avg	Cmp.	Red.	Con.	Avg	Cpl.	Den.	Rea.	
Initial Retrieval	16.30	14.67	17.18	17.05	60.91	18.29	71.57	92.88	32.08	46.59	20.05	29.60	36.43
BGE-Reranker$_{L.}$	20.30	18.89	21.32	20.70	63.05	28.04	67.86	93.26	38.34	54.35	21.03	39.64	40.57
MonoT5 (3B)	19.90	19.00	20.35	20.36	62.76	26.92	68.74	92.61	38.17	53.97	21.29	39.26	40.28
RankT5 (3B)	20.12	19.47	20.43	20.47	62.86	26.86	68.61	93.12	37.97	54.08	21.17	38.67	40.32
RankLlama (7B)	19.63	19.13	19.94	19.81	63.08	26.07	69.51	93.66	37.38	53.46	20.75	37.94	40.03
Rearank (7B)	19.28	18.47	19.93	19.44	63.34	26.24	70.23	93.55	38.00	53.83	20.99	39.17	40.21
ReasonRank (7B)	19.23	18.59	19.88	19.23	64.99	27.90	72.52	94.54	38.82	54.62	21.39	40.45	41.01
Rank4Gen (8B)	25.85	25.07	26.08	26.39	67.65	28.04	80.41	94.49	38.12	54.01	21.99	38.37	43.87
SetR (8B)	33.44	29.98	35.57	34.77	64.04	27.82	71.41	92.88	40.08	55.35	24.62	40.28	45.85
RubricRanker (8B)	36.15	33.69	**38.07**	36.69	63.31	28.25	65.10	96.58	40.96	56.25	24.97	**41.65**	46.80
AdaTutoRank (8B)	**36.59**	**39.55**	33.25	**36.98**	**69.28**	**28.72**	**81.49**	**97.64**	**41.04**	**56.93**	**25.32**	40.88	**48.97**

and nearly reproduces "Only Distillation" (47.79), so neither signal can dominate. ❹ ***Privileged Information.*** Restricting the hint to one fixed form degrades performance, yielding 47.48 (rubrics), 46.88 (sibling sets), and 47.62 (reflections). The extremes fail oppositely: abstract rubrics give a badly wrong rollout no foothold, while a sibling-set overrides the equally valid choices of a strong one. The adaptive tutoring hint of Eq. 8 is therefore necessary.

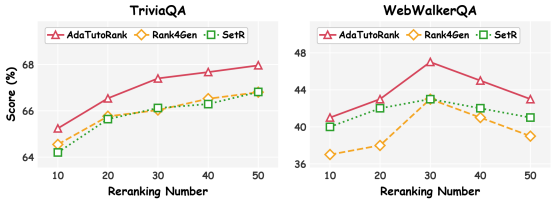


Figure 4: Results of different candidate doc numbers.

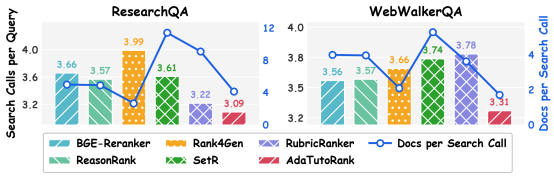


Figure 5: Search behaviour of different rerankers.

Table 3: Ablation Study of AdaTutoRank on the RAG and Deep Research Benchmarks.

Setting	NQ	PopQA	HealthB	RQA	Overall
AdaTutoRank	36.87	38.52	48.57	72.78	**49.19**
Training Stage					
w/o ATO	36.45	38.25	40.96	69.67	46.33
w/o SFT	34.79	38.45	45.24	69.67	47.04
Advantage Superposition					
Only RL	31.88	36.23	46.11	66.06	45.07
Only Distillation	36.43	38.39	46.38	69.95	47.79
Interpolation Coefficient					
$\lambda = 0.9$	33.16	36.17	47.42	69.71	46.61
$\lambda = 0.5$	36.70	38.47	45.82	70.00	47.75
Privileged Information					
Only Rubrics h^*	36.51	38.43	47.20	67.78	47.48
Only Reflection $\bar{h}$	35.60	37.77	47.36	69.77	47.62
Only Sibling-Set $\bar{h}$	35.82	38.22	43.11	70.36	46.88

Varying Number of Candidate Documents and Search Call Rounds. We further examine how AdaTutoRank behaves as its input and downstream usage vary: Figure 4 reports performance across candidate pool sizes, and Figure 5 inspects the retrieval behavior it induces in a deep research agent. ❶ ***AdaTutoRank is robust to the candidate pool size.*** In Figure 4, AdaTutoRank stays strongest as the pool grows from 10 to 50. On TriviaQA, performance rises steadily with pool size; on WebWalkerQA, all three rerankers peak at 30 and decline thereafter, implicating the retrieval source rather than the reranker: web results for agent-issued sub-queries contain a long tail of low-value pages that inflate the context without adding evidence. ❷ ***AdaTutoRank is more efficient in search agents.*** As shown in Figure 5, AdaTutoRank incurs the fewest search calls on the two deep research benchmarks while selecting fewer documents per round. Its sets are thus both sufficient and compact: the agent meets its information need without issuing extra queries.

5 Conclusion and Discussion

We present AdaTutoRank, a setwise reranker for RAG and deep research that optimizes set-level utility. To make utility supervisable, we represent it as hierarchical query-specific rubrics spanning the document, set, and global levels, which score individual documents, their interaction, and the set as a whole. Our two-stage framework cold-starts the policy on rubric-conditioned silver labels, then runs adaptive tutoring optimization, superposing the outcome reward with a token-level distillation advantage from a hint matched to each rollout's quality. At inference it selects a high-quality subset directly, without query-specific rubrics or hints. Experiments show consistent gains at both the answer and setwise levels, with fewer search calls in deep research agents. More broadly, rubrics can supply not only evaluation but also dense token-level supervision distilled back into the policy.

AI USE STATEMENT

In this work, we used generative AI tools for data construction, reward modeling, evaluation, and grammar polishing, as detailed below. We did not use generative AI tools for research ideation, experimental design, or theoretical derivation, and the remaining disclosure categories are not applicable to this work. We have reviewed all AI-assisted content and take responsibility for the final text, claims, and artifacts of this paper.

- **Usage 1: Training data construction.** As described in Section 3.2 and 3.3, DeepSeek-V4-Pro is used to instantiate the query-specific rubrics from each query and its reference answer, and to produce the silver labels for cold-start supervised fine-tuning.
- **Usage 2: Rubric-based reward.** As described in Section 3.4, DeepSeek-V4-Flash serves as the rubric-based judge that scores a selected document set during training, a practice consistent with current work on LLM-as-a-judge reward models.
- **Usage 3: Evaluation.** As described in Section 4.1, the open-ended deep research benchmarks and the setwise-level benchmark are scored by LLM judges, following the protocol released with each benchmark.
- **Usage 4: Paper grammar polishing.** The paper was drafted by the authors and subsequently polished for grammar and wording with LLM assistance, a practice consistent with current research norms.

ETHICS STATEMENT

AdaTutoRank aims to improve the quality of the evidence that large language models consume. Our study uses only public question-answering datasets and publicly accessible web search results, involves no human subjects, and collects no personally identifiable information. Since web content may be inaccurate or biased and the selected set conditions downstream generation, we discourage deployment in high-stakes settings without human oversight.

REPRODUCIBILITY STATEMENT

To guarantee reproducibility, we report full training details and hyperparameters in Appendix C, and will release all source code, data, and trained weights (*e.g.,* on Hugging Face) post-review to aid the community in verifying and reproducing our findings.

REFERENCES

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. Healthbench: Evaluating large language models towards improved human health. *ArXiv*, abs/2505.08775, 2025. 7, 16

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'Arcy, David Wadden, Matt Latzke, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke S. Zettlemoyer, Graham Neubig, Dan Weld, Doug Downey, Wen tau Yih, Pang Wei Koh, and Hanna Hajishirzi. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *ArXiv*, abs/2411.14199, 2024. 5, 19

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations*, volume 2024, pp. 35549–35562, 2024. 19

Jingwen Deng, Jihao Huang, Zhen Hao Wong, Hao Liang, Quanqing Xu, Bin Cui, and Wentao Zhang. Data-centric perspectives on agentic retrieval-augmented generation: A survey. In *Annual Meeting of the Association for Computational Linguistics*, 2026. 3

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Licheng Zhang, Xiaorui Wang, and Zhendong Mao. Deepresearch bench: A comprehensive benchmark for deep research agents. In *International Conference on Learning Representations*, volume 2026, pp. 42414–42448, 2026. 7, 16

Yongqi Fan, Yuxiang Chu, Zhentao Xia, Xiaoyang Chen, Jie Liu, Haijin Liang, Jin Ma, Ben He, Yingfei Sun, Dezhi Ye, and Tong Ruan. Rank4gen: Rag-preference-aligned document set selection and ranking. *ArXiv*, abs/2601.11273, 2026. 3, 7, 17

Yunfan Gao, Yun Xiong, Xinyu Gao, Kang-Xiang Jia, Jin Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qian-Yu Guo, Meng Wang, and Hao-Fen Wang. Retrieval-augmented generation for large language models: A survey. *ArXiv*, abs/2312.10997, 2023. 1, 3

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020. 5, 7, 16, 18

Jonas Hubotter, Frederike Lubeck, Lejs Deen Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. *ArXiv*, abs/2601.20802, 2026. 2

Kailin Jiang, Lei Liu, Jian Xi, Hui Xu, Junlin Liu, Baochen Fu, Bin Li, Yu Lu, Haibo Shi, et al. Beyond relevance-centric retrieval: Rubric-oriented document set selection and ranking. *arXiv preprint arXiv:2607.19747*, 2026. 2, 4, 7, 16

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025. 1

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017. 7, 15

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 6769–6781, 2020. 8

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc V. Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019. 5, 7, 15, 18

Dahyun Lee, Yongrae Jo, Haeju Park, and Moontae Lee. Shifting from ranking to set selection for retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17606–17619, 2025. 3, 7, 17

Tongyi Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, Kuan Li, Liangcai Su, Litu Ou, Liwen Zhang, Pengjun Xie, Rui Ye, Wenbiao Yin, Xinmiao Yu, Xinyu Wang, Xixi Wu, Xuanzhong Chen, Yida Zhao, Zhen Zhang, Zhengwei Tao, Zhongwang Zhang, Zile Qiao, Chenxi Wang, Donglei Yu, Gang Fu, Haiyang Shen, Jiayi Yang, Jun Lin, Junkai Zhang, Kuijie Zeng, Li Yang, Hailong Yin, Maojia Song, Mingshi Yan, Peng Xia, Qian Xiao, Rui Min, Rui Ding, Runnan Fang, Shaowei Chen, Shen Huang, Shihang Wang, Shihao Cai, Weizhou Shen, Xiaobin Wang, Xin Guan, Xinyu Geng, Yi Shi, Yuning Wu, Zhuo Chen, Zijian Li, and Yong Jiang. Tongyi deepresearch technical report. 2025. 1, 3

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2023. 1

Wenhan Liu, Yu Lu, Qiaolin Xia, Hui Xu, Tong Zhao, Jian Xi, Yutao Zhu, Haijin Liang, Haibo Shi, Hao Wang, et al. Training documents reranker with search rubrics for deep research agent. *arXiv preprint arXiv:2608.03527*, 2026a. 2, 3, 4, 7, 17, 19

Wenhan Liu, Xinyu Ma, Weiwei Sun, Yutao Zhu, Yuchen Li, Dawei Yin, and Zhicheng Dou. Reasonrank: Empowering passage ranking with strong reasoning ability. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22007–22031, 2026b. 3, 7, 17

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. Fine-tuning llama for multi-stage text retrieval. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023. 3, 7, 17

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9802–9822, 2023. 7, 16

Mihran Miroyan, Tsung-Han Wu, Logan King, Tianle Li, Jiayi Pan, Xinyan Hu, Wei-Lin Chiang, Anastasios Angelopoulos, Narges Norouzi, Joseph E Gonzalez, et al. Search arena: Analyzing search-augmented llms. In *International Conference on Learning Representations*, volume 2026, pp. 41109–41146, 2026. 5, 19

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. Document ranking with a pre-trained sequence-to-sequence model. In *Findings of the association for computational linguistics: EMNLP 2020*, pp. 708–718, 2020. 3, 7, 16

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. Rankvicuna: Zero-shot listwise document reranking with open-source large language models. *arXiv preprint arXiv:2309.15088*, 2023a. 3, 7, 17

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. Rankzephyr: Effective and robust zero-shot listwise reranking is a breeze! *arXiv preprint arXiv:2312.02724*, 2023b. 3, 7, 17

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, 2023. 7, 16

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 3505–3506, 2020. 19

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and trends® in information retrieval*, 4(1-2):1–174, 2009. 7

Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xin-Ran Zhao, Molly Park, Samuel G. Finlayson, David Sontag, Tyler Murray, Sewon Min, Pradeep Dasigi, Luca Soldaini, Faeze Brahman, Wen tau Yih, Tongshuang Wu, Luke S. Zettlemoyer, Yoon Kim, Hanna Hajishirzi, and Pang Wei Koh. Dr tulu: Reinforcement learning with evolving rubrics for deep research. *ArXiv*, abs/2511.19399, 2025. 1, 4, 17

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Jun-Mei Song, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *ArXiv*, abs/2402.03300, 2024. 6

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*, 2026. 2

Zhengliang Shi, Yiqun Chen, Haitao Li, Weiwei Sun, Shiyu Ni, Yougang Lyu, Run-Ze Fan, Bowen Jin, Yi-Xuan Weng, Minjun Zhu, Qiujie Xie, Xinyu Guo, Qu Yang, Jiayi Wu, Jujia Zhao, Xia-Qiang Tang, Xinbei Ma, Cunxiang Wang, Jiaxin Mao, Qingyao Ai, Jen-Tse Huang, Wenxuan Wang, Yue Zhang, Yiming Yang, Zhaopeng Tu, and Zhaochun Ren. Deep research: A systematic survey. *ArXiv*, abs/2512.02038, 2025. 1, 3

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. 5, 18

Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, et al. Webwalker: Benchmarking llms in web traversal. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10290–10305, 2025. 5, 7, 16, 19

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian yun Nie. C-pack: Packed resources for general chinese embeddings. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023. 1, 3, 7, 16

Renjun Xu and Jingwen Peng. A comprehensive survey of deep research: Systems, methodologies, and applications. *arXiv preprint arXiv:2506.12594*, 2025. 1

An Yang, An-Feng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bo-Wen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Jingren Zhou, Junyan Lin, Kai Dang, Keqin Bao, Ke-Pei Yang, Le Yu, Li-Chun Deng, Mei Li, Min Xue, Ming-Ze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tian-Yi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Ying-Er Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhen-Ru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. 2025. 8, 19

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing*, 2018. 5, 18

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *ArXiv*, abs/2210.03629, 2022. 1, 3, 4

Tian-Zhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *ArXiv*, abs/2602.12275, 2026. 2

Li S Yifei, Allen Chang, Chaitanya Malaviya, and Mark Yatskar. Researchqa: Evaluating scholarly question answering at scale across 75 fields with survey-mined questions and rubrics. *Transactions of the Association for Computational Linguistics*, 14:1344–1368, 2026. 7, 16

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *International Conference on Learning Representations*, volume 2024, pp. 29862–29883, 2024. 1

Le Zhang, Bo Wang, Xipeng Qiu, Siva Reddy, and Aishwarya Agrawal. Rearank: Reasoning re-ranking agent via reinforcement learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2458–2471, 2025. 3, 7, 17

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *ArXiv*, abs/2601.18734, 2026. 2, 14, 19

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 3: system demonstrations)*, pp. 400–410, 2024. 19

Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. Rankt5: Fine-tuning t5 for text ranking with ranking losses. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022. 1, 7, 16

A What Adaptive Tutoring Optimization Optimizes

This section identifies the objective behind Eq. 11, shows that the distillation branch already contains a reverse-KL anchor to the cold start, and explains why the teacher is frozen at θ_{sft} rather than tracking the policy.

Setup. Write $x = (q, \mathcal{C})$ for the hint-free context and $x_h = (q, \mathcal{C}, \boldsymbol{h}_i)$ for the hint-augmented one. For a rollout $y_i \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$ and a position ℓ, abbreviate the per-token conditionals on the prefix $y_{i,<\ell}$ as p_θ, p_{old}, and p_ϕ^h for $\pi_\theta(\cdot \mid x, \cdot)$, $\pi_{\theta_{\text{old}}}(\cdot \mid x, \cdot)$, and $\pi_\phi(\cdot \mid x_h, \cdot)$, and let $p_\phi := \pi_\phi(\cdot \mid x, \cdot)$ be the same frozen teacher evaluated *without* the hint, a quantity the algorithm never computes but which is well defined and needed below. Every divergence that follows is summed over positions with prefixes drawn from $\pi_{\theta_{\text{old}}}$, the form the token-level advantage implements. With this notation Eq. 9 reads $A_{i,\ell}^{\text{ATD}} = \log p_\phi^h(\hat{y}_{i,\ell}) - \log p_{\text{old}}(\hat{y}_{i,\ell})$ with $\hat{y}_{i,\ell} \sim p_{\text{old}}$, and Eq. 10 reads $A_{i,\ell}^{\text{ATO}} = \lambda A_i^{\text{GRPO}} + (1-\lambda) m_i A_{i,\ell}^{\text{ATD}}$, where $m_i = \mathbb{1}[r_i \neq -1]$ is the format mask of Algorithm 1. It excludes malformed rollouts from the distillation branch, so a rollout receiving $r_i = -1$ under Eq. 6 is supervised by the outcome advantage alone and the hint that Eq. 8 assigns to it never takes effect; every statement below is therefore vacuous for such rollouts.

Assumption 1 (Detached advantages at the snapshot)**.** *Both advantages are evaluated at parameters fixed within the update, so they are constants with respect to θ, and $\epsilon > 0$. Gradients are taken at $\theta = \theta_{\text{old}}$, which is exact for the first inner step of each update.*

Lemma 1 (The clip is inactive at the snapshot)**.** *Under Assumption 1,* $\nabla_\theta \mathcal{L}_{\text{policy}}\big|_{\theta_{\text{old}}} = -\mathbb{E}\big[A_{i,\ell}^{\text{ATO}} \nabla_\theta \log p_\theta(\hat{y}_{i,\ell})\big]$.

Proof. At $\theta = \theta_{\text{old}}$ we have $\rho_{i,\ell} = 1$, interior to $[1-\epsilon, 1+\epsilon]$, so in a neighbourhood both arguments of the minimum equal $\rho_{i,\ell} A_{i,\ell}^{\text{ATO}}$. Differentiating and using $\nabla_\theta \rho_{i,\ell} = \nabla_\theta \log p_\theta$ at $\rho_{i,\ell} = 1$ gives the claim. □

Lemma 2 (The snapshot divergence is first-order flat)**.** $\nabla_\theta D_{\text{KL}}(p_\theta \,\|\, p_{\text{old}})\big|_{\theta=\theta_{\text{old}}} = 0$.

Proof. Differentiating $D_{\text{KL}}(p_\theta \| p_{\text{old}}) = \mathbb{E}_{p_\theta}[\log p_\theta - \log p_{\text{old}}]$ gives $\mathbb{E}_{p_\theta}[\nabla_\theta \log p_\theta (\log p_\theta - \log p_{\text{old}})] + \mathbb{E}_{p_\theta}[\nabla_\theta \log p_\theta]$. At $\theta = \theta_{\text{old}}$ the first bracket vanishes pointwise and the second term is zero since $\mathbb{E}_{p_\theta}[\nabla_\theta \log p_\theta] = \nabla_\theta \sum_y p_\theta(y) = 0$. □

Proposition 1 (ATO is reinforcement learning with a reverse-KL pull toward the hint-conditioned teacher)**.** *Under Assumption 1, the first inner step of an update follows the ascent direction of*

$$J_{\text{ATO}}(\theta) \;=\; \lambda\, J_{\text{GRPO}}(\theta) \;-\; (1-\lambda)\, m_i\, D_{\text{KL}}\big(p_\theta \,\big\|\, p_\phi^h\big), \qquad J_{\text{GRPO}}(\theta) := \mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\big[A^{\text{GRPO}}(y)\big]. \tag{13}$$

Proof. Let $J_{\text{ATD}}(\theta) := \mathbb{E}_{\hat{y}\sim p_\theta}[\log p_\phi^h(\hat{y}) - \log p_{\text{old}}(\hat{y})]$. Adding and subtracting $\mathbb{E}_{p_\theta}[\log p_\theta]$ gives $J_{\text{ATD}} = -D_{\text{KL}}(p_\theta \| p_\phi^h) + D_{\text{KL}}(p_\theta \| p_{\text{old}})$, and Lemma 2 kills the gradient of the second summand, so $\nabla_\theta J_{\text{ATD}}\big|_{\theta_{\text{old}}} = -\nabla_\theta D_{\text{KL}}(p_\theta \| p_\phi^h)$. By Lemma 1 and the score-function identity $\nabla_\theta \mathbb{E}_{p_\theta}[f] = \mathbb{E}_{p_\theta}[f \nabla_\theta \log p_\theta]$, valid for any θ-independent f and applied to A^{GRPO} at the level of rollouts and to $A_{i,\ell}^{\text{ATD}}$ at the level of tokens, $\nabla_\theta \mathcal{L}_{\text{policy}} = -\lambda \nabla_\theta J_{\text{GRPO}} - (1-\lambda) m_i \nabla_\theta J_{\text{ATD}}$. □

Eq. 13 makes the role of λ precise: it trades the outcome objective against a regularizer whose target is selected per rollout by Eq. 8. Fixing $\boldsymbol{h}_i \equiv h^*$ recovers self-distillation from a single privileged context (Zhao et al., 2026) as the special case in which that target does not depend on i.

Proposition 2 (The distillation branch contains an anchor to the cold start)**.** *Let the* hint lift *be $\Delta(y) := \log p_\phi^h(y) - \log p_\phi(y)$, which depends only on the frozen ϕ and is therefore independent of θ. Then $D_{\text{KL}}(p_\theta \| p_\phi^h) = D_{\text{KL}}(p_\theta \| p_\phi) - \mathbb{E}_{p_\theta}[\Delta]$, and hence, with $\phi = \theta_{\text{sft}}$,*

$$J_{\text{ATO}}(\theta) \;=\; \underbrace{\lambda\, J_{\text{GRPO}}(\theta)}_{\text{outcome objective}} \;+\; \underbrace{(1-\lambda) m_i\, \mathbb{E}_{p_\theta}\big[\Delta\big]}_{\text{token-level hint shaping}} \;-\; \underbrace{(1-\lambda) m_i\, D_{\text{KL}}\big(p_\theta \,\big\|\, \pi_{\theta_{\text{sft}}}\big)}_{\text{reverse-KL anchor to the cold start}}. \tag{14}$$

Proof. Both divergences share the term $\mathbb{E}_{p_\theta}[\log p_\theta]$, so their difference is $\mathbb{E}_{p_\theta}[\log p_\phi - \log p_\phi^h] = -\mathbb{E}_{p_\theta}[\Delta]$. Substituting into Eq. 13 and using $p_\phi = \pi_{\theta_{\text{sft}}}(\cdot \mid x, y_{i,<\ell})$ gives the claim. □

Corollary 1 (An explicit KL penalty rescales a coefficient rather than adding a constraint)**.** *Since Δ is the only h_i-dependent factor in Eq. 14 while the anchor is common to all rollouts, averaging over a group leaves the anchor with coefficient $(1-\lambda)\bar{m}$, where $\bar{m} = \frac{1}{G}\sum_i m_i$. An explicit penalty $\beta D_{\text{KL}}(p_\theta \| \pi_{\theta_{\text{sft}}})$ enters in the same position and with the same sign, so it changes that coefficient to $(1-\lambda)\bar{m} + \beta$ and supplies no regularizer of a different kind.*

This is the precise sense of the claim in Section 3.4 that freezing the teacher lets us omit the KL penalty; the anchor also weakens in proportion to the share of malformed rollouts, which $m_i = 0$ excludes. We do not claim that either mechanism bounds the cumulative displacement from θ_{sft}, only that both are first-order regularizers of the same type.

Why the teacher is frozen. A natural alternative is to let the teacher track the policy and to restore stability with an explicit KL term. Two results argue against it.

Proposition 3 (A moving self-teacher annihilates the anchor)**.** *Substituting $\phi := \theta_{\text{old}}$ into Proposition 2 gives $p_\phi = p_{\text{old}}$, so the anchor becomes $D_{\text{KL}}(p_\theta \| p_{\text{old}})$, whose gradient vanishes at $\theta = \theta_{\text{old}}$ by Lemma 2. The surviving objective $\lambda J_{\text{GRPO}} + (1-\lambda) m_i \mathbb{E}_{p_\theta}[\Delta^{\text{old}}]$, with $\Delta^{\text{old}}(y) := \log \pi_{\theta_{\text{old}}}(y \mid x_h) - \log p_{\text{old}}(y)$, contains no divergence to any parameter setting held fixed across updates.*

An explicit KL term is therefore not an optional addition in that design but a repair of what the substitution removed, and it is the term Eq. 14 already provides.

Proposition 4 (Hint-conditioned behaviour is never supervised)**.** *Neither stage places a gradient on the policy's hint-conditioned conditionals. The cold-start objective of Eq. 3 is defined on hint-free inputs, since R_q is withheld, and by Lemma 1 the ATO gradient depends on θ only through $\nabla_\theta \log \pi_\theta(\hat{y}_{i,\ell} \mid x, y_{i,<\ell})$. Hence for every hint the map $\boldsymbol{h} \mapsto \pi_\theta(\cdot \mid x_h)$ appears in no objective and evolves only as a side effect of updating shared parameters.*

This is the decisive difference between the two designs. Exploiting a hint is an in-context ability inherited from pre-training: it is used during training, never trained, and never invoked at inference. A self-teacher therefore draws its targets from an object that drifts with the policy while no term maintains its quality, so Δ^{old} degrades in an uncontrolled way. It also admits a shortcut, since a policy that becomes insensitive to hints makes $\Delta^{\text{old}} \equiv 0$ and zeroes the distillation gradient regardless of the utility of its selected sets, at no cost under hint-free inference. Freezing $\phi = \theta_{\text{sft}}$ removes both problems: Δ is a fixed function of y, and since p_ϕ^h does not depend on θ, the distillation gradient can vanish only at stationary points of $D_{\text{KL}}(p_\theta \| p_\phi^h)$. It is also cheaper, requiring one extra forward pass per update rather than two, because $\pi_{\theta_{\text{old}}}(\cdot \mid x)$ is already computed for the ratio of Eq. 12.

Limitations. Three caveats delimit the above. The identities are first-order at $\theta = \theta_{\text{old}}$ and exact only for the first inner step, after which the clip may activate. The hint lift Δ is measured under ϕ rather than under the current policy, so it estimates the effect of the hint off-policy; this is the price of a stationary target. Finally, Eq. 14 is a penalty rather than a constraint, and neither it nor an explicit KL term bounds the cumulative displacement of π_θ from $\pi_{\theta_{\text{sft}}}$, since the clipped objective enforces only a per-update trust region.

B Details of Experimental Setup

B.1 Details of RAG, Deep Research and Setwise Benchmarks

We evaluate AdaTutoRank on ten benchmarks: five short-form RAG datasets, four deep research datasets, and one setwise-level benchmark. Brief descriptions and evaluation protocols are given below.

RAG Benchmarks. All five datasets require short, closed-ended answers and are scored by exact match (EM) against the gold answers, with no LLM judge involved. ❶ **Natural Questions** (Kwiatkowski et al., 2019) consists of real queries issued to Google search, paired with answers annotated from Wikipedia. Its questions reflect authentic information needs and are mostly single-hop, making it a standard testbed for open-domain retrieval-augmented QA. ❷ **TriviaQA** (Joshi et al.,

2017) contains over 95K trivia-style question-answer pairs, each accompanied by independently collected evidence documents. Because the questions are written without reference to a specific passage, answering them requires retrieving evidence rather than matching surface lexical cues. ❸ **PopQA** (Mallen et al., 2023) comprises 14K entity-centric questions generated from Wikidata triples, with a long-tail distribution of entity popularity. It is designed to expose the limits of parametric knowledge, and thus isolates the contribution of the retrieved evidence. ❹ **2WikiMultihopQA** (Ho et al., 2020) is constructed from Wikipedia and Wikidata with approximately 192K samples, providing explicit evidence paths in the form of triples and covering four question types: comparison, inference, compositional, and bridge-comparison. ❺ **Bamboogle** (Press et al., 2023) is a curated test set of 125 two-hop questions designed so that no single search query can retrieve the answer, requiring models to identify and use intermediate bridging entities.

Deep Research Benchmarks. All four datasets require long-form, open-ended responses and are scored by LLM judges following the protocol released with each benchmark. For comparability and efficiency, we evaluate 100 queries per benchmark, sampling randomly where the official set is larger. ❻ **WebWalkerQA** (Wu et al., 2025) focuses on complex web and website-level question answering. Solving its questions requires agents to search, browse, and connect information distributed across web pages, rather than relying on a single retrieved passage. ❼ **HealthBench** (Arora et al., 2025) is a healthcare-oriented benchmark that evaluates whether models can answer medically related user questions in a helpful, accurate, and safe manner. Its queries often require careful interpretation of user intent and reliable synthesis of evidence, making it suitable for evaluating long-form responses in high-stakes domains. ❽ **DeepResearchBench** (Du et al., 2026) evaluates general-domain deep research capabilities using open-ended questions, emphasizing comprehensiveness, depth of analysis, instruction following, and readability. Its 100 queries are evenly split between English and Chinese; we score the generated articles on the four criteria above and report the macro average. Following the official protocol (Du et al., 2026), the Jina API is used to scrape URLs for evidence snippets when needed, while for our system outputs we use the URL contents collected by the corresponding search or browsing tools. ❾ **ResearchQA** (Yifei et al., 2026) contains research-style questions that require collecting external evidence and producing synthesized long-form answers. It tests whether an agent can identify useful information, organize evidence, and provide a coherent response to complex information needs. We use its official 100-question subset for evaluating deep research systems and report the average rubric score.

Setwise-level Benchmark. The remaining benchmark scores the retrieved set itself rather than the answer produced from it. ❿ **SetwiseEvalKit** (Jiang et al., 2026) instantiates query-specific criteria and scores a set along nine dimensions organized at three levels: the document level (relevance, authenticity, quality), the set level (complementarity, redundancy, conflict), and the global level (completeness, density, reachability). All metrics are reported such that higher is better, and we report the per-level averages together with the overall score. Our main results use the short-form scenario, where each query is paired with a candidate pool from which a set must be selected. Note that the benchmark instantiates its own criteria with its own judge, independently of the rubrics and reward judge used in our training pipeline (Section 3.4); we therefore treat the setwise-level results as a diagnostic that localizes where the answer-level gains come from, rather than as an independent replication of them.

B.2 Details of Reranker

We compare AdaTutoRank with 11 rerankers spanning three categories, together with an unreranked lower bound. Each baseline is described below.

❶ **Adhoc reranking.** These methods score candidates by pointwise relevance to the query and return a ranked list, from which the top-5 documents are taken as the selected set. **BGE-Reranker-Large** (Xiao et al., 2023) is a cross-encoder released as part of the C-Pack suite. Query and document are concatenated into a single sequence so that full bidirectional attention can be applied across them, and the resulting representation is mapped to a relevance score. We adopt the `bge-reranker-large` checkpoint, an XLM-RoBERTa model of roughly 560M parameters. **MonoT5** (Nogueira et al., 2020) recasts relevance estimation as sequence-to-sequence generation: conditioned on a query–document pair, the model emits either "true" or "false", and the normalized probability assigned to "true" is used as the score. **RankT5** (Zhuang et al., 2022) also builds on T5 but reads a scalar directly off the decoder logits instead of going through token probabilities, which allows it to be trained with

pointwise, pairwise, or listwise ranking losses. We adopt the `rankt5-base` checkpoint fine-tuned on MS MARCO. **RankLLaMA** (Ma et al., 2023) fine-tunes LLaMA-2-7B for pointwise reranking. The hidden state of the final token serves as the sequence representation and is projected linearly to a scalar, with training driven by a pairwise ranking loss over hard negatives mined from both sparse and dense retrievers. **RankVicuna** (Pradeep et al., 2023a) is an early fully open listwise reranker for zero-shot settings. Starting from Vicuna-7B, it is trained by permutation distillation against a GPT-3.5-based teacher, learning to emit a reordered permutation of the candidates within a sliding window. **RankZephyr** (Pradeep et al., 2023b) is a 7B listwise reranker built on Zephyr. It combines permutation distillation from GPT-3.5 and GPT-4 with direct preference optimization, narrowing the gap to proprietary rerankers while remaining openly available.

❷ **Reasoning-enhanced reranking.** These methods expose an explicit reasoning process before committing to a ranking decision, trading inference cost for accuracy on harder queries. **Rearank** (Zhang et al., 2025) treats listwise reranking as a reasoning task and optimizes the reasoning-then-ranking behavior with reinforcement learning, so that competitive rankings can be obtained from a modest amount of labeled supervision. **ReasonRank** (Liu et al., 2026b) targets reasoning-intensive retrieval. It is first warmed up on synthesized reasoning traces for ranking and then refined with a ranking-oriented reward, which yields substantial gains on benchmarks where relevance cannot be judged from surface overlap.

❸ **Setwise reranking.** Rather than producing a full ordering, these methods predict the retained subset directly, and therefore differ in what supervision they use to define a good set. **Rank4Gen** (Fan et al., 2026) selects a subset with the downstream generator in mind, using the quality of the answer produced from a candidate set as the training signal, so that selection is optimized for generation rather than for human-perceived relevance. **SetR** (Lee et al., 2025) learns set selection by imitating a strong teacher, distilling the teacher's subset decisions into a smaller policy that can assemble a set in a single pass. **RubricRanker** (Liu et al., 2026a) is the closest baseline to our work. It scores a candidate set against query-specific rubrics and uses the aggregated score as a reinforcement learning reward, which supervises set composition but delivers a single set-level scalar shared by all its documents.

B.3 Details of Generator and Search Agent

RAG Generator. In the RAG scenario, the selected documents are consumed by *Llama-3.1-8B-Instruct*, which conditions on the top-k passages returned by each reranker and produces a short answer. A mid-scale instruction-tuned generator is deliberately chosen here: it is strong enough to exploit good evidence, yet not so strong that it can compensate for a poor set from parametric knowledge alone, so the differences we observe downstream can be attributed to the quality of the selected set rather than to the generator.

Deep Research Agent. In the deep research scenario, we build on *DR.Tulu-8B* (Shao et al., 2025), an open-source deep research agent initialized from Qwen3-8B and trained end-to-end with reinforcement learning against evolving rubrics. The agent runs an autonomous multi-turn loop that alternates between internal planning (`think`), tool invocation (`call_tool`), and answer synthesis (`answer`), with web search, page browsing, and paper retrieval available as tools. We keep the agent fixed and swap only the reranker inside its retrieval pipeline, so that any change in the final report can be traced to how the evidence at each search step was composed.

C Implementation Details

C.1 Details of Meta Rubrics

The meta rubrics are dimension-level templates rather than criteria themselves: for each query they are instantiated into concrete, checkable criteria, each carrying an importance weight. The nine dimensions are organized into three levels according to what they examine.

Level 1: Document-level. These dimensions examine each document in isolation and are rated per document and averaged over the set. ❶ **Relevance (Rel.)** asks whether a document addresses the information need actually expressed by the query, as opposed to merely sharing surface terms with it. ❷ **Authenticity (Aut.)** asks whether the factual claims a document makes are consistent with the

reference answer, penalizing outdated, misattributed, or fabricated statements. It is the only dimension requiring the reference answer, and is therefore available during training but never at inference. ❸ **Quality (Qua.)** asks whether the target information can be extracted with little effort, rewarding clear structure and self-contained statements while penalizing fragmented or navigation-heavy text.

Level 2: Set-level. These dimensions examine how documents interact and are rated once for the whole set. None of them is a property of any single document, which is why a pointwise reranker cannot observe them. ❹ **Complementarity (Cmp.)** asks whether the members contribute distinct information elements, so that the set is worth more than any one of them. ❺ **Redundancy (Red.)** asks whether multiple documents convey the same information without incremental contribution. It is not the inverse of complementarity: a set may avoid redundancy yet still fail to be complementary, if its members are distinct but jointly uninformative. ❻ **Conflict (Con.)** asks whether documents make contradictory statements about the same fact, which forces the consumer to adjudicate between sources.

Level 3: Global-level. These dimensions evaluate the set in its role as the input context of a downstream model, and are likewise rated once for the whole set. ❼ **Completeness (Cpl.)** asks whether the set covers every key element a satisfactory answer requires. Unlike complementarity, which compares members against one another, completeness compares the set against the information need. ❽ **Density (Den.)** asks whether useful content dominates the context rather than irrelevant filler, thereby capturing the trade-off between broader coverage and a concise, undiluted input. ❾ **Reachability (Rea.)** asks whether a model with no external knowledge could arrive at the correct answer from this set alone, and is thus the dimension most directly aligned with the answer-level metrics.

Instantiation. Given the meta rubrics above, a query q, and its reference answer a, DeepSeek-V4-Pro instantiates one or more criteria under each dimension. Each criterion must be phrased as a question about the document set and must name specific entities, facts, or figures from q and a; vague formulations such as "contains relevant content" are rejected. The resulting criteria and weights form R_q, which drives both the silver labels and the reward of Eq. 5. Full prompts are in Appendix F.

Inapplicable criteria. Several dimensions presuppose a configuration that a given set may not exhibit: conflict needs two comparable claims about the same fact, and complementarity needs two relevant documents carrying different required points. Scoring such a criterion on a 0–10 scale would reward a set for trivially "not conflicting" when nothing could conflict in the first place. The judge therefore returns `null` rather than a score whenever a criterion's applicability prerequisite is unmet, and every `null` criterion is dropped from both the numerator and the denominator of Eq. 5, so the weights renormalize over the applicable criteria alone and an inapplicable dimension neither raises nor lowers the reward. Relevance is never `null`, and completeness and reachability take 0 rather than `null` when the relevant subset covers nothing, so at least one criterion always survives and $u(\mathcal{S} \mid q)$ stays well defined.

C.2 Details of Training Query

Table 4: **Training Data Composition.** Number of queries used in each training stage, grouped by scenario. OSch: OpenScholar; SA_s/SA_l: SearchArena short-/long-form; Glaive: GlaiveAI-Reasoning-v1; WW: WebWalker.

Stage	RAG					Deep Research						Total
	Sum	HotpotQA	NQ	2Wiki	MuSiQue	Sum	OSch	SA_s	Glaive	WW	SA_l	
SFT	3,433	1,644	1,209	432	148	5,525	2,439	1,176	831	739	340	8,958
ATO	1,065	511	426	83	45	2,492	1,252	527	276	306	131	3,557
Total	**4,498**	2,155	1,635	515	193	**8,017**	3,691	1,703	1,107	1,045	471	**12,515**

Training queries are drawn from both scenarios the reranker is meant to serve, since the two differ in what a query looks like.

RAG queries. We sample user questions from the training splits of four multi-hop question answering datasets, HotpotQA (Yang et al., 2018), Natural Questions (Kwiatkowski et al., 2019), 2WikiMultihopQA (Ho et al., 2020), and MuSiQue (Trivedi et al., 2022). These questions are short and

closed-ended, and each comes with a gold answer, which is required because our rubric generation conditions on the reference answer.

Deep research queries. In this scenario a reranker is invoked on the sub-queries an agent issues to itself, not on the original user question, and such sub-queries are rarely released with open-source datasets. Rather than regenerate them with an agent of our own, which would tie the training distribution to one particular agent, we reuse the agent-issued sub-queries released by RubricRanker (Liu et al., 2026a). They originate from OpenScholar (Asai et al., 2024), SearchArena (Miroyan et al., 2026), GlaiveAI-Reasoning-v1-20M, and WebWalker-Silver (Wu et al., 2025), and therefore span scientific, general-purpose, reasoning-oriented, and web-navigation information needs.

Candidate pools and filtering. For every retained query we build a candidate pool with the retriever of the corresponding scenario, and sample the pool size from 10 to 40 per query so that the policy is exposed to varying numbers of candidates. Queries for which rubric generation fails to produce a well-formed set of weighted criteria are discarded. Table 4 reports the resulting number of queries per source for the cold-start and the adaptive tutoring optimization stages.

Combining the two sources yields queries that differ in form, length, and difficulty, which in turn yields more diverse rubrics and, empirically, better generalization across the two scenarios.

C.3 Cold-Start SFT Stage

Silver label construction. Silver labels are produced by DeepSeek-V4-Pro, which is given the query, the candidate pool, and the query-specific rubrics, and returns the identifiers of the documents to retain. The two scenarios use different prompt templates, since an agent-issued sub-query carries an accompanying query intent that a user question does not.

Fine-tuning setup. We fine-tune Qwen3-8B (Yang et al., 2025) on these labels with LlamaFactory (Zheng et al., 2024) on 8 NVIDIA H20 GPUs. The model input follows the same scenario-specific format used for label construction, comprising the query and the candidate documents, plus the query intent in the deep research case, but excluding the query-specific rubrics: they depend on a reference answer that is unavailable at inference, so withholding them keeps the training and inference inputs identical. Optimization uses a learning rate of `5e-6`, a per-device batch size of 1, and 8 gradient accumulation steps, giving an effective batch size of 64, for one epoch under BF16 mixed precision. Throughput is improved with DeepSpeed ZeRO-2 (Rasley et al., 2020) and FlashAttention-2 (Dao, 2024).

C.4 Adaptive Tutoring Optimization Stage

Implementation. The second stage is implemented in the VERL framework on top of its on-policy self-distillation recipe (Zhao et al., 2026), which we extend from a single fixed privileged context to the reward-conditioned hint space of Eq. 8. Rollouts are generated with SGLang under a tensor-parallel degree of 8. Each update combines the outcome-driven GRPO advantage with the token-level distillation advantage of Eq. 9 in the convex form of Eq. 10, so that a single policy-gradient step is taken on the combined advantage. The two terms are not on a common scale: the GRPO advantage is standardized within each rollout group, whereas the distillation advantage is a raw log-probability gap. We therefore read λ as a mixing weight tuned empirically, not as a literal ratio of gradient contributions, and select it by the ablation in Table 3 rather than by matching magnitudes. Following the reverse-KL instantiation of on-policy self-distillation, the distillation advantage is computed as the log-probability gap between the hint-conditioned teacher and the rollout snapshot on the sampled tokens, with the teacher distribution truncated to its top 128 logits. The teacher is held at the cold-start checkpoint θ_{sft} throughout the stage and is never updated, which also supplies the implicit anchor that lets us disable the explicit KL penalty. Rollouts whose output violates the required identifier format are excluded from the distillation branch and supervised by the outcome reward alone, so that malformed sequences are never used as distillation targets.

Hint assignment. Each rollout is routed to one of three hints by its rubric reward r_i, with bucket boundaries at $\tau_2 = 3$ and $\tau_1 = 6$ on the judge's 0–10 scale: rollouts scoring at least 6 receive the rubrics, those in $(3, 6)$ receive a corrective reflection, and those scoring at most 3 receive the sibling set. The two hints that build on the sibling set are additionally gated in relative mode, i.e. they are

Algorithm 1: AdaTutoRank: Adaptive Tutoring Optimization

Require : Cold-start checkpoint θ_{sft}; query set $\mathcal{Q}$ with candidate pools $\mathcal{C}$ and rubrics R_q; judge J; self-selector $\mathcal{S}$; self-reflector $\mathcal{F}$; thresholds $\tau_1 > \tau_2$; group size G; convex weight λ; clip parameter ϵ; learning rate η

Ensure : Hint-free reranking policy π_θ

$\theta \leftarrow \theta_{\text{sft}},\ \phi \leftarrow \theta_{\text{sft}}$ *// teacher π_ϕ stays frozen throughout*

for *each policy update* **do**

  $\theta_{\text{old}} \leftarrow \theta$; sample a query batch $\mathcal{B} \subset \mathcal{Q}$

  / On-policy rollouts, rubric rewards, and outcome advantages */*

  foreach *query* $q \in \mathcal{B}$ **do**

    Sample $\mathcal{G}_q = \{y_i\}_{i=1}^{G}$ with $y_i \sim \pi_{\theta_{\text{old}}}(\cdot \mid q, \mathcal{C})$

    foreach *rollout* $y_i \in \mathcal{G}_q$ **do**

      if y_i *violates the identifier format* **then**

        $r_i \leftarrow -1,\ m_i \leftarrow 0$ *// excluded from distillation*

      else

        $\mathcal{S}_i \leftarrow \text{Parse}(y_i),\ r_i \leftarrow u(\mathcal{S}_i \mid q)$ via J and $R_q,\ m_i \leftarrow 1$

      end

    end

    $A_i^{\text{GRPO}} \leftarrow \big(r_i - \text{mean}(\{r_j\})\big)/\text{std}(\{r_j\})$ for all i

    / Sibling set from the self-selector, scored by the same judge */*

    $\bar{h}_q \leftarrow \mathcal{S}\big(q, \mathcal{C}, R_q; \pi_{\theta_{\text{old}}}\big),\ r_{\bar{h}} \leftarrow u(\bar{h}_q \mid q)$

    / Reward-conditioned hint assignment with admission gate */*

    foreach *rollout* $y_i \in \mathcal{G}_q$ **do**

      if $r_i \geq \tau_1$ **then**

        $\boldsymbol{h}_i \leftarrow h^* = R_q$ *// strong rollout: criteria only*

      else if $r_{\bar{h}} \leq r_i$ **then**

        $\boldsymbol{h}_i \leftarrow h^* = R_q$ *// gate rejects: fall back*

      else if $\tau_2 < r_i < \tau_1$ **then**

        $\boldsymbol{h}_i \leftarrow \hat{h}_i = \mathcal{F}\big(q, \mathcal{C}, y_i, \bar{h}_q; \pi_{\theta_{\text{old}}}\big)$ *// corrective reflection*

      else

        $\boldsymbol{h}_i \leftarrow \bar{h}_q$ *// weak rollout ($r_i \leq \tau_2$): sibling set*

      end

    end

  end

  / Paired re-scoring and joint optimization */*

  foreach *minibatch of sampled tokens* (i, ℓ) **do**

    $A_{i,\ell}^{\text{ATD}} \leftarrow \text{sg}\big[\log \pi_\phi(\hat{y}_{i,\ell} \mid q, \mathcal{C}, \boldsymbol{h}_i, y_{i,<\ell}) - \log \pi_{\theta_{\text{old}}}(\hat{y}_{i,\ell} \mid q, \mathcal{C}, y_{i,<\ell})\big]$ *// no regeneration*

    $A_{i,\ell} \leftarrow \lambda A_i^{\text{GRPO}} + (1-\lambda)\, m_i A_{i,\ell}^{\text{ATD}}$

    $\rho_{i,\ell}(\theta) \leftarrow \pi_\theta(\hat{y}_{i,\ell} \mid q, \mathcal{C}, y_{i,<\ell}) \big/ \pi_{\theta_{\text{old}}}(\hat{y}_{i,\ell} \mid q, \mathcal{C}, y_{i,<\ell})$

    $\mathcal{L}(\theta) \leftarrow -\mathbb{E}_{i,\ell}\big[\min\big(\rho_{i,\ell} A_{i,\ell},\ \text{clip}(\rho_{i,\ell}, 1-\epsilon, 1+\epsilon) A_{i,\ell}\big)\big]$

    $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}(\theta)$ *// no explicit KL; π_ϕ anchors*

  end

end

admitted only when the sibling set outscores the rollout it supervises ($r_{\bar{h}} > r_i$, with zero margin), and otherwise the rollout falls back to the rubric hint.

Reward service. Rubric rewards are produced by a locally deployed DeepSeek-V4-Flash judge, served in FP8 with SGLang behind an internal endpoint, which avoids the cost and rate limits of a commercial API and keeps the reward model fixed across all runs. To reduce the number of judge calls, all set- and global-level criteria of a rollout are scored in a single request that returns one score per criterion, and document-level criteria are likewise batched per document. Requests are issued by a parallel reward manager with 512 concurrent workers.

Table 5: Representative adaptive tutoring hints for two training queries, one from the RAG pool and one from the deep research pool. For each query we show three rollouts of the same policy that fall into different reward bands. $\hat{y}_i$ is the subset selected by that rollout itself, and r_i its judge reward; the hint is chosen according to r_i and is shown only to the teacher. The rubric hint states the criteria only, the corrective reflection names the documents to add or drop, and the sibling set gives a concrete reference subset.

Hint type	r_i	**Rollout** $\hat{y}_i$	**Privileged information to the teacher**
RAG. Do both films *The Garden of Eden* (1998) and *One Night in Yoshiwara* have directors that share the same nationality?			
Rubrics	9.25	[2] [14]	[doc-level] *Relevance (weight 5): does each document directly discuss the director (by name) and nationality of The Garden of Eden (1998) and/or One Night in Yoshiwara?* [set-level] *Complementarity (weight 5): does the set contain at least one document identifying the director and nationality of each of the two films, so that together they enable the comparison?* *. . . seven further dimensions across the three tiers.*
Corrective reflection	4.00	[1] [2]	*The chosen documents [1] and [2] are insufficient. [1] is irrelevant, and [2] only covers The Garden of Eden, giving its director Alessandro D'Alatri and his Italian nationality. Document [14] supplies the director of One Night in Yoshiwara, Max Ophüls, and his French nationality. Only [2] and [14] are needed.*
Sibling-set	3.67	[1] [2]	*[2] [14]* (scored 9.25 by the judge)
Deep Research. Illuvium Pre-season 2 rewards 10k ILV			
Rubrics	8.84	[1] [3]	[doc-level] *Relevance (weight 5): does each document address the Pre-season 2 reward allocation in terms of the total ILV amount, rather than merely mentioning Illuvium or Pre-season 2 without the pool size?* [set-level] *Conflict (weight 5): are all statements about the Pre-season 2 ILV amount mutually consistent, with no document claiming a different total (e.g. 5,000 or 15,000 ILV)?* *. . . seven further dimensions across the three tiers.*
Corrective reflection	4.83	[1] [6]	*Document [3] should be included: it states that "an additional 10,000 ILV awaits in Pre-Season 2" and comes from the official Illuvium account. Document [6] should be excluded: it refers to 10,000 ILV added to an existing 20,000 ILV pool, which is ambiguous and conflicts with the figure in [3].*
Sibling-set	0.00	[2] [6]	*[1] [3]* (scored 9.32 by the judge)

Hyperparameters. Training runs on 32 NVIDIA H20 GPUs (4 nodes $\times$ 8 GPUs). We use a training batch of 64 prompts per step with a PPO mini-batch of 64 and 8 rollouts per prompt, a learning rate of `5e-6` with 10 warmup steps and a cosine schedule decaying to 0.1 of its peak, a maximum prompt length of 16,000 tokens, and the interpolation coefficient $\lambda = 0.7$ selected by the ablation in Table 3. The maximum response length is 64 tokens, since the policy emits the bracketed identifier sequence of Eq. 6 directly, with no reasoning trace before the selection. Queries are shuffled once per epoch and drawn without replacement, and the trailing incomplete batch is dropped, so one epoch amounts to $\lfloor 3{,}557/64 \rfloor = 55$ steps over the ATO queries of Table 4, leaving 37 queries unused in that epoch. Validation runs every 5 steps and checkpointing every 15. All results we report come from the checkpoint at step 30, which attains the best validation reward; it has therefore been optimized on $30 \times 64 = 1{,}920$ distinct queries, slightly over half of a single epoch, so the reported policy is obtained well before the training queries are exhausted.

C.5 Algorithm and Hint Examples

Algorithm 1 summarizes the adaptive tutoring loop. At each update, the frozen snapshot $\pi_{\theta_{\text{old}}}$ serves three roles: it samples the rollout group for every query, it assembles the sibling set under the query-specific rubrics, and it contrasts each rollout against that sibling set to produce a corrective

Table 6: **Setwise-Level Performance Comparison on SetwiseEvalKit (Long-form Scenario).** The **best** (1st) and second-best (2nd) results are highlighted. The per-level **Avg** is the arithmetic mean of the three dimensions within that level, and **Overall** is the mean of all nine dimensions. All metrics are reported such that higher is better (↑).

Ranker	Doc-Level				Set-Level				Global-Level				Overall
	Avg	Rel.	Aut.	Qua.	Avg	Cmp.	Red.	Con.	Avg	Cpl.	Den.	Rea.	
Initial Retrieval	20.20	29.28	11.38	19.93	64.52	31.90	71.06	90.61	23.51	22.26	31.93	16.35	36.08
BGE-Reranker$_{L.}$	21.57	31.66	12.76	20.30	62.10	33.33	65.09	87.87	25.27	23.48	33.96	18.37	36.31
MonoT5 (3B)	21.76	32.23	12.57	20.47	63.85	34.70	69.61	87.25	25.11	24.89	32.80	17.64	36.91
RankT5 (3B)	20.01	29.85	11.34	18.84	62.31	30.83	69.62	86.49	23.52	22.06	32.54	15.97	35.28
RankLlama (7B)	20.96	29.95	13.32	19.62	62.69	32.53	64.76	90.77	23.93	22.58	31.38	17.82	35.86
Rearank (7B)	22.50	32.63	13.24	21.64	63.64	34.37	67.53	89.02	25.41	24.41	33.64	18.18	37.18
ReasonRank (7B)	22.95	33.94	13.39	21.52	**64.82**	36.94	67.98	89.55	26.80	25.92	35.34	19.14	38.19
Rank4Gen (8B)	23.57	35.09	13.45	22.18	62.54	28.00	**74.64**	84.98	23.45	20.34	36.48	13.53	36.52
SetR (8B)	21.13	30.46	12.45	20.48	63.31	32.53	65.73	**91.68**	24.10	24.38	29.86	18.06	36.18
RubricRanker (8B)	29.70	34.42	23.90	30.78	60.26	**41.65**	53.75	85.39	**33.66**	**35.13**	33.57	**32.27**	41.21
AdaTutoRank (8B)	**34.10**	**40.97**	**26.00**	**35.32**	63.27	39.09	62.78	87.94	33.17	32.00	**37.71**	29.79	**43.51**

Table 7: Ablation Study of AdaTutoRank on the RAG and Deep Research Benchmarks.

Setting	RAG					Deep Research				Overall
	NQ	Triv	PopQA	2Wiki	Bambo	WebW	HealthB	DRB	RQA	
AdaTutoRank	36.87	66.54	38.52	32.52	20.80	41.00	48.57	49.89	72.78	**45.28**
Training Stage										
w/o ATO	36.45	66.72	38.25	33.25	20.80	38.00	40.96	48.61	69.67	43.63
w/o SFT	34.79	66.14	38.45	33.23	14.40	37.00	45.24	48.92	69.67	43.09
Advantage Superposition										
Only RL	31.88	64.28	36.23	26.49	17.60	43.00	46.11	48.64	66.06	42.25
Only Distillation	36.43	66.34	38.39	32.85	16.80	46.00	46.38	49.18	69.95	44.70
Interpolation Coefficient										
$\lambda = 0.9$	33.16	65.86	36.17	28.94	18.40	41.00	47.42	48.40	69.71	43.23
$\lambda = 0.5$	36.70	66.74	38.47	32.74	17.60	39.00	45.82	48.74	70.00	43.98
Privileged Information										
Only Rubrics h^*	36.51	66.86	38.43	32.28	16.00	39.00	47.20	48.41	67.78	43.61
Only Reflection $\hat{h}$	35.60	66.30	37.77	31.35	21.60	47.00	47.36	48.74	69.77	45.05
Only Sibling-Set $\bar{h}$	35.82	66.65	38.22	32.40	16.00	43.00	43.11	47.72	70.36	43.70

reflection. The rubric judge then scores both the rollouts and the sibling set, and each rollout is routed to one of the three hints by its own reward, falling back to the rubrics whenever the sibling set fails the admission gate. The hint-conditioned teacher π_ϕ, held at the cold-start checkpoint θ_{sft}, re-scores the sampled tokens under the hint-augmented context, while the same snapshot re-scores them under the ordinary context, so the resulting advantage is constant in θ and gradients reach the policy only through the importance ratio of Eq. 12. The resulting ATD advantage is combined convexly with the GRPO advantage and optimized in a single clipped policy-gradient step, after which the updated policy becomes the next snapshot.

Table 5 provides representative hints of the three types, drawn from rollouts in the high, middle, and low reward bands of the same queries.

D More Experimental Results

Beyond the results reported in the main text, this section supplements four groups of experiments. Appendix D.1 extends the setwise-level evaluation from the short-form to the long-form scenario, verifying that the gains in evidence quality are not specific to short closed-ended queries. Appendix D.2 gives the per-benchmark breakdown of all four ablations, whose overall scores are averaged over the nine answer-level benchmarks rather than the subset shown in the main text. Appendix D.3 completes the two analyses of Section 4.2 on the benchmarks omitted there, covering candidate pool sizes on NQ and PopQA and search behavior on HealthBench and DRB. Appendix D.4 then turns

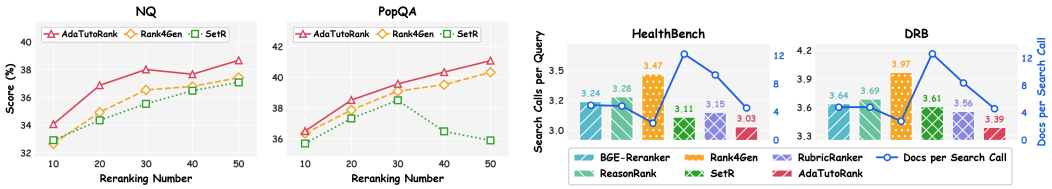


Figure 6: **Left:** performance across candidate pool sizes on NQ and PopQA. **Right:** search calls per query and documents per search call on HealthBench and DRB.

from outcomes to the training process itself, tracking how the three hint types are distributed, how often the admission gate admits them, and how strong the sibling-set reference remains as the policy improves, so as to verify that the adaptive assignment is genuinely exercised rather than collapsing into a single form.

D.1 Setwise-Level Evaluation

Table 6 extends the setwise evaluation to the long-form scenario, where the same pattern holds. AdaTutoRank attains the best overall score of 43.51, ahead of the strongest baseline RubricRanker by 2.30 points. It leads on every document-level metric, raising the document-level average from 29.70 to 34.10, and also improves the set-level average over RubricRanker (63.27 vs. 60.26), so the gains over the closest baseline appear at both the level of individual documents and the level of their interaction. At the global level it leads on density (37.71), indicating that the selected context stays informative rather than padded, but trails RubricRanker on completeness (32.00 vs. 35.13) and reachability (29.79 vs. 32.27), and hence on the global-level average (33.17 vs. 33.66); its long-form advantage therefore comes from document- and set-level composition rather than from global coverage. Taken together, the improvement comes from how documents are chosen and combined rather than from retaining more of them.

D.2 Complete Ablation Experiment Results

Table 7 reports the per-benchmark breakdown of all four ablations, averaged over the nine benchmarks. Removing either training stage costs a comparable amount, 1.65 points without ATO and 2.19 without the cold start, so neither stage is dispensable; the cold start matters most on the short-form benchmarks, where its removal drops Bamboogle from 20.80 to 14.40, consistent with its role as a stable initialization for the format the policy must emit. Bamboogle alone contributes 0.71 of that 2.19, and over the remaining eight benchmarks the two stages cost 1.85 and 1.66 respectively, matching the ordering obtained on the four-benchmark subset of Table 3.

For the advantage design, the outcome signal alone is the weakest configuration overall (42.25) and degrades most sharply on the multi-hop sets, falling from 32.52 to 26.49 on 2Wiki, whereas the distillation signal alone already reaches 44.70; superposing the two recovers the best score on both scenario groups. The interpolation coefficient degrades on both sides of $\lambda = 0.7$, and each end drifts toward the corresponding single-signal regime, 43.23 at $\lambda = 0.9$ against 42.25 for the outcome signal alone and 43.98 at $\lambda = 0.5$ against 44.70 for the distillation signal alone. Finally, every fixed hint form underperforms the adaptive assignment, and the breakdown shows why no single form suffices: among the three fixed-hint variants, rubrics alone are the strongest on NQ (36.51) yet tie with the sibling set for the weakest on Bamboogle (16.00), while reflections alone reverse this ordering (35.60 on NQ against 21.60 on Bamboogle), so the form that helps one query type is not the form that helps another.

D.3 Varying Number of Candidate Documents and Search Call Rounds

Figure 6 reports the two analyses on the remaining benchmarks. On NQ and PopQA, AdaTutoRank stays the strongest reranker at every pool size and improves monotonically as the pool grows from 10 to 50, whereas SetR peaks at 30 candidates on PopQA and then declines, indicating that a larger pool benefits our policy rather than distracting it. Since the training pools are sampled from 10 to 40 (Appendix C.2), the continued gain at 50 lies outside the range the policy was trained on, so set selection transfers to pools larger than any it saw during training rather than being tuned to a

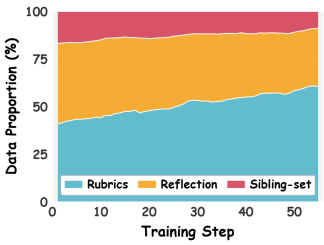


Figure 7: Proportion of rollouts per hint type over training.

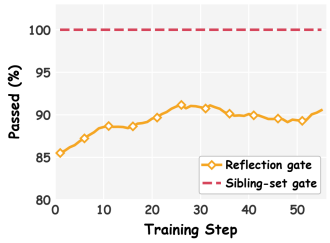


Figure 8: Fraction of rollouts passing the quality gate.

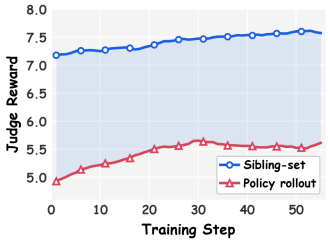


Figure 9: Judge reward of the sibling-set and rollouts.

particular pool size. On HealthBench and DRB, it issues the fewest search calls per query (3.03 and 3.39) while also returning the fewest documents per call, so the agent reaches its answer with both fewer rounds and a smaller context.

D.4 Training Dynamics of Adaptive Tutoring Hints

Since the sibling set is produced by the policy's own snapshot under rubric guidance, one may ask whether the three-way assignment is genuinely exercised during training or silently collapses into the rubric hint. Figure 7 shows that all three forms remain in use throughout: the rubric share grows from roughly two fifths of the rollouts to about three fifths, the sibling-set share shrinks from under a fifth to about a tenth, and reflections make up the remainder, easing from about two fifths to roughly three tenths. The shift is in the expected direction, as an improving policy places more of its rollouts in the high-reward band and needs the most prescriptive form less often, so the hint space behaves like a curriculum that fades rather than a fixed context. Figure 8 shows that the admission gate is rarely the binding constraint: it passes for every rollout in the sibling-set band and for 85% to 90% of those in the reflection band, so fallbacks are the exception. Figure 9 explains why. The sibling set is scored between 7.2 and 7.6 by the judge while the policy's own rollouts move from 5.0 to 5.6, so rubric-guided re-selection extracts a set that stays about two points ahead of what the policy produces by default, and the reference remains informative even as the policy strengthens.

E Case Study

E.1 Case Study of Query-Specific Rubrics

To make the query-specific rubrics concrete, Tables 8 and 9 show the full rubric set generated for one RAG query and one deep research query. The RAG case is a bridging multi-hop question whose rubrics pin down a single verifiable fact and the two-document evidence chain needed to reach it, whereas the deep research case is an open-ended mechanistic query whose rubrics instead enumerate the distinct aspects a useful subset must jointly cover. Both are instantiated over the same three tiers and nine dimensions, but the weights and the content of each dimension are tailored to the query: relevance and completeness dominate in the factoid case, while complementarity across documents carries the most weight in the deep research case. This contrast illustrates why a fixed, query-agnostic criterion is insufficient for set-level selection.

E.2 Case Study of RAG

Tables 10, 11 and 12 trace three queries end to end, and two patterns recur. First, the passages that actually establish the answer sit *low* in the retrieval order: they are ranked 2 and 19 on NQ, 17 and 20 on 2WikiMultihopQA, and 7 and 12 on PopQA, while the head of the list is occupied by passages that share the query terms without answering it. PopQA shows this most sharply, where passage 3 contains the string *RnB* but describes a Vietnamese singer who admires Houston. Ordering candidates by relevance therefore does not surface the evidence a generator needs, which is exactly the mismatch our objective targets. Second, retrieving the right passage is not sufficient. On NQ, RankT5, SetR and RubricRanker all forward both [2] and [19], yet each pairs them with passages dating other generations of the machine and the generator still answers *October 2012*; on PopQA, three baselines reach [7] and [12] but keep the novelist alongside and answer *Romantic Novels*. AdaTutoRank returns

Query: In which year did the country that encompassed Beyra become independent?

Level	Category	Description	Weight
Document-level	Relevance	Does each document directly discuss which country Beyra belongs to, or the independence date/year of that country, rather than being about unrelated topics that merely mention the word "Beyra" or the year "1960"?	5
Document-level	Authenticity	For each document that states the year in which the country encompassing Beyra became independent, is the stated year consistent with the objective fact that it is 1960, without misstating it as a different year?	5
Document-level	Quality	Is each document clearly structured and focused, allowing a reader to directly extract the independence year (1960) and, if present, the identity of the country that encompasses Beyra, rather than burying these key facts in noisy or irrelevant content?	2
Set-level	Complementarity	Does the document set contain, across its documents, at least one that specifies which country encompasses Beyra AND at least one that provides the independence year of that country, so that the two pieces of information together enable determining the correct year?	4
Set-level	Redundancy	Does the document set avoid having two or more documents that convey essentially the same information — that the country containing Beyra gained independence in 1960 — with no incremental contribution (e.g., no additional detail, source, or nuance)?	2
Set-level	Conflict	Across the documents, are all statements about the independence year of the country that encompassed Beyra free of mutual contradiction — specifically, do they all support the year 1960, without any document claiming a different year such as 1958 or 1962?	4
Global-level	Completeness	Does the whole document set provide both (a) enough information to identify which country encompasses Beyra, and (b) the independence year of that country, leaving no gap that would prevent a direct answer of "1960"?	5
Global-level	Density	Across the documents, does the text that directly helps identify the country containing Beyra or states its independence year make up the majority of each document's length, rather than each document being dominated by unrelated content with only a small fraction mentioning Beyra or the year 1960?	2
Global-level	Reachability	Using only the document set and no external knowledge, can the full reasoning chain be completed — determining which country encompasses Beyra, then obtaining that country's independence year — to produce the correct answer "1960"?	4

Table 8: Case study of query-specific rubrics for a multi-hop question in the RAG scenario.

exactly the two supporting passages in all three cases and is the only method that answers all three correctly, showing that discarding a conflicting document matters as much as including a useful one.

E.3 Case Study of Set Selection

Table 13 inspects a single query under the full rubric. The two passages that establish the answer are [1] and [12], and both selectors retain them; the difference is that RubricRanker additionally forwards [2] and [3], which describe CHBC's own disaffiliation without naming CFJC-TV. These extras are free riders: the set stays complete and the answer stays reachable, so the two selectors receive an identical global-level score of 3.33. What the extras do change is the other two levels. Document-level criteria are averaged over the set, so two passages that fail relevance and authenticity pull both scores from 4.0 down to 2.5, and since the extras contribute no distinct information, complementarity falls from 2.0 to 1.0. The overall score separates the two sets (3.22 against 2.67) only because utility is scored at the document and set levels as well as globally; a judgement made on the final answer alone would have rated them equally.

Query: endothelial senescence preeclampsia telomere p16 p21 oxidative stress sFlt-1 biomarkers

Level	Category	Description	Weight
Document-level	Relevance	Does each document directly address the mechanistic relationship between preeclampsia, later-life cardiovascular disease (CVD), and endothelial senescence (including pathways such as oxidative stress, sFlt-1, senescence markers p16/p21, and the SASP), rather than merely mentioning preeclampsia or CVD without discussing the endothelial senescence connection?	5
Document-level	Authenticity	For each document, does its stated risk magnitude for future cardiovascular disease after preeclampsia align with the established range of approximately 2–4-fold increased risk (of hypertension, ischemic heart disease, stroke, heart failure), rather than claiming a dramatically different number or denying the association?	4
Document-level	Authenticity	For each document, does its description of sFlt-1 correctly state that it is a soluble VEGF receptor-1 that binds and neutralizes VEGF and PlGF, acting as an anti-angiogenic factor that promotes endothelial dysfunction and senescence, rather than misrepresenting its function or directional effect?	4
Document-level	Quality	Is each document clearly structured and focused enough that a reader can directly extract the relationships between preeclampsia, endothelial senescence mechanisms (oxidative stress, sFlt-1, p16/p21, SASP), and later cardiovascular disease risk, without these elements being buried in large amounts of unrelated obstetrics or vascular biology content?	2
Set-level	Complementarity	Does the document set collectively cover all the key information elements required to fully explain the relationship — preeclampsia as a state of exaggerated vascular aging, evidence for oxidative stress-induced cellular senescence (p16/p21, telomere dysfunction), the role of sFlt-1 and VEGF deprivation, the SASP-driven inflammatory milieu, the connection to later CVD risk, and biomarker implications — such that the documents jointly build a complete picture?	5
Set-level	Redundancy	Does the document set avoid having two or more documents that convey essentially the same core information — for example, multiple documents merely restating in similar words that preeclampsia involves oxidative stress and endothelial dysfunction without adding distinct aspects like sFlt-1 signaling, p16/p21 biomarker data, or long-term CVD risk evidence?	2
Set-level	Conflict	Are the documents in the set free of mutual contradictions on key factual elements — for example, not having one document claim a 5-fold CVD risk while another says no increased risk, or contradicting on sFlt-1's role (one stating it inhibits angiogenesis and another claiming it promotes angiogenesis) — so that the set consistently supports the correct mechanistic account?	4
Global-level	Completeness	Does the whole document set, taken as a single unit, cover every essential information element needed to completely explain the relationship between preeclampsia, endothelial senescence, and cardiovascular disease, including preeclampsia pathophysiology, cellular senescence mechanisms, evidence for endothelial senescence in PE, persistent senescence linking to future CVD, and biomarker candidates?	5
Global-level	Density	Across the documents in the set, does the text that actually addresses the preeclampsia–endothelial senescence–CVD relationship make up the majority of each document's length, rather than each document being dominated by irrelevant content such as general obstetrics, unrelated vascular biology, or filler text where only a small snippet is useful?	2
Global-level	Reachability	Using only the document set and no external knowledge, can a reader trace the full reasoning chain — from preeclampsia's placental oxidative stress and anti-angiogenic state, to induction of endothelial senescence via DNA damage, p53/p21/p16 pathways, telomere dysfunction and SASP, and then to the 2–4-fold increased risk of later cardiovascular disease — with every step grounded in the documents?	4

Table 9: Case study of query-specific rubrics for an agent query in the deep research scenario.

Benchmark: Natural Questions **Query:** when did macbook pro 13 inch come out? **Gold answer:** *October 2008*

ID	Candidate document
[1]	MacBook Pros come with ExpressCard/34 slots, which replace the PC Card slots found in the PowerBook G4. All first generation 15-inch models have two USB 2.0 ports and one FireWire 400 port, while the 17-inch …
[2]	quoted at eight hours, with 80 percent of this charge remaining after 1,000 charge-discharge cycles. At Apple's Worldwide Developers Conference (WWDC) on June 8, 2009, it was announced that the 13-inch unibody MacBook would be upgraded and re-branded as a MacBook Pro, leaving only the white polycarbonate MacBook in the MacBook line. [. . .]
[3]	Pro follows the design of the previous two generations with an all-metal unibody enclosure and separated black keys. A few apparent design changes are a thinner chassis, thinner screen bezel, larger trackpad, …
[4]	the 15-inch version. All models come with 4 GB of system memory that is upgradeable to 8 GB. Battery life was also extended further in this update, to an estimated ten hours for the 13-inch and 8–9 hours …
[5]	i9 MacBook Pro was slower than the 2017 MacBook Pro and stated, "This isn't a problem with Intel's Core i9, it's Apple's thermal solution." When Lee put the i9 MacBook Pro inside a freezer, the render …
[6]	MacBook Pro The MacBook Pro (sometimes abbreviated as MBP) is a line of Macintosh portable computers introduced in January 2006 by Apple Inc. It is the high-end model of the MacBook family and is currently …
[7]	FireWire 800 port and all except the 17-inch models would receive an SD card slot. The 17-inch model would retain its ExpressCard/34 slot. For the 13-inch MacBook Pro, the Kensington lock slot was moved …
[8]	13-inch model now comes with a 128GB storage option, down from the base 256GB storage. On July 12, 2018 Apple updated the Touch Bar models with Intel Coffee Lake quad-core processors in 13-inch models …
[9]	MacBook Pro is Apple's higher end laptop available in both 13-inch and 15-inch configurations. A redesigned MacBook Pro was introduced on October 27, 2016, which is thinner and lighter than the previous …
[10]	or red if the battery is charging. MagSafe can be found on the MacBook, MacBook Pro and MacBook Air notebook computers, as well as the Apple LED Cinema Display. The MacBook and the 13-inch MacBook Pro use …
[11]	MacBook Pro. None of the three 17-inch models of the MacBook Pro have used any pentalobe screws. The MacBook Air has seen more extensive use of pentalobe screws than the MacBook Pro. All five versions …
[12]	its eight-hour capacity. Some sources even reported up to eight hours of battery life for the 13- and 15-inch MacBook Pros during casual use, while others reported around six hours. Like the 17-inch …
[13]	ports, and the default RAM on premium models was increased to 8 GB. Following this announcement, the 17-inch model was discontinued. After a media event on October 22, 2013 Apple discontinued all second …
[14]	mistaken for 5-point Torx screws. This was the only internal usage of pentalobe screws; all following MacBook Pros use the "Tri-Wing" security bit to attach the battery to the internal frame, or else …
[15]	Pro units without Touch Bar, manufactured between October 2016 and October 2017, may have the built-in battery expanded, which is also known as "swelling". Apple initiated a free replacement program …
[16]	macOS High Sierra 10.13.4. Devices using HDMI, previous generation Thunderbolt, and USB will require an adapter to connect to the MacBook Pro. The models come with a 3.5 mm headphone jack, although …
[17]	added as an option for the 17-inch model. Processors were updated to "Penryn" cores, which are built on the 45 nanometer process (65 nanometer "Merom" cores were previously used), and hard drive …
[18]	is thinner than its predecessor and is the first to include a high-resolution Retina Display. A 13-inch variant was released in October 2012. The fourth generation MacBook Pro was announced on October …
[19]	processors later that year. The product's second iteration, known as the "unibody" model, has a casing made from a single piece of aluminum. It debuted in October 2008 in 13- and 15-inch screen sizes. In January 2009, the 17-inch model was updated with the same unibody design. [. . .]
[20]	requirements lists the following requirements for Mac OS X Lion and Mac OS X Mountain Lion: Apple lists the following requirements for Mac OS X 10.5 Leopard and Mac OS X 10.6 Snow Leopard: Officially …

Reranker	Selected documents	Answer	Reranker	Selected documents	Answer
Initial Retrieval	[1] [2] [3] [4] [5]	2006	RankZephyr (7B)	[6] [18] [19] [7] [2]	October 2012
BGE-Reranker$_{L.}$	[9] [6] [18] [8] [7]	October 2012	Rearank (7B)	[6] [2] [18] [9] [19]	October 2012
MonoT5 (3B)	[18] [9] [6] [13] [2]	October 2012	ReasonRank (7B)	[9] [2] [17] [18] [6]	July 2020
RankT5 (3B)	[6] [9] [18] [2] [19]	October 2012	Rank4Gen (8B)	[18] [2] [9]	October 2012
RankLlama (7B)	[18] [13] [9] [7] [6]	January 2006	SetR (8B)	[2] [6] [18] [19]	October 2012
RankVicuna (7B)	[2] [9] [8] [12] [13]	–	RubricRanker (8B)	[18] [2] [19] [12] [7]	October 2012
AdaTutoRank (8B)	[2] [19]	**October 2008**			

Table 10: Case study on Natural Questions for the query *when did macbook pro 13 inch come out?* The upper block lists the 20 BM25 candidates, where blue marks the spans that establish the answer. The lower block lists the documents each method feeds to the generator (the top-5 for a ranking method, the chosen subset for a set selector), where green marks a selected passage that supports the answer and red one that does not, together with the answer the generator then produced. Passages 6, 9 and 18 date other generations of the machine, January 2006, October 2016 and October 2012; eleven of the twelve baselines forward at least one of them and nine answer with one of those years. Ours selects exactly the two passages that date the 13-inch model. [. . .] marks an elision inside a passage.

Benchmark: 2WikiMultihopQA **Query:** Where did Dolores Hope's husband die? **Gold answer:** *Toluca Lake, Los Angeles*

ID	Candidate document
[1]	and her daughter. Vera suggests Dolores take advantage of the up-coming eclipse to solve her problem, singing, "accidents can be an unhappy woman's best friend." The second act opens with Dolores preparing food and liquor …
[2]	Honorary Board Member of the humanitarian organization Wings of Hope. On May 29, 2003, Dolores was at her husband's side as he celebrated his 100th birthday; he died two months later on July 27, 2003. They had been married …
[3]	accounts to fund their escape, she discovers Joe has stolen everything she had saved. In despair, she breaks down crying at work, forcing her to confide her troubles in Vera. An unusually sympathetic Vera reveals she has had …
[4]	Terrill was around. Toto mentions that Terrill was with Dolores the whole time. The chief then interrogates Dolores about Terrill and where her husband was. She mentions that she couldn't help but fall for Terrill and she …
[5]	receives a letter from the residents of Ramsdale, who have learnt that Dolores has gone missing and are pressing for answers. Knowing that his situation is precarious and contemplating what to do, Humbert eventually receives …
[6]	Lo's Diary Lo's Diary () is a 1995 novel () by Pia Pera, retelling Vladimir Nabokov's novel "Lolita" from the point of view of "Dolores Haze (Lolita)". It depicts Dolores as a sadist and a controller of everyone around her; …
[7]	involved in a centuries-old battle between the Animus and the Dolore clans. Long ago one of a pair of a Book of Curses was stolen by the Dolore, and because of that they were put under a curse where they become beasts and …
[8]	as Meri Bell), played integral roles in encouraging former First Lady Betty Ford to establish the Betty Ford Center. Bell was Ford's sponsor in Alcoholics Anonymous. Ford's obituary in USA Today related developments after …
[9]	wealthy, elderly employer, Vera Donovan. The novel is presented as a transcript of her statement, told to the local constable and a stenographer. Dolores wants to make clear to the police that she did not kill Vera, whom she …
[10]	the British Empire Exhibition, where the exotic foreign displays intrigued him, or possibly through his friend Matthew Smith. In 1925 Epstein invited Sunita, Enver and Anita to live at his home at Guilford Street in London …
[11]	of the day off by Vera. Dolores and Selena had an argument about Dolores' suspicions regarding Joe's sexual abuse. Selena fled home for the weekend to work at a hotel, where guests had flocked for the eclipse. Joe soon …
[12]	Dolores Richard Spikes Dolores Margaret Richard Spikes (August 24, 1936 – June 1, 2015) was an American mathematician and university administrator. Born in Baton Rouge, Dolores Richard attended public and parochial schools …
[13]	1970, the Spanish right-wing newspaper "ABC" reported that the PCE and the Kremlin had reached a new pact whereby the Spanish party dropped its censure of the Soviet invasion of Czechoslovakia in exchange for the Kremlin's …
[14]	kill her. Vera dies and the police begin a murder investigation. Dolores' daughter, Selena St. George, is a successful journalist, living in New York City, who battles depression and substance abuse. Selena arrives in town …
[15]	that he created a trade union at the 'bodega' where he worked during the war. Dona Dolores: Dona Dolores is more conservative than her liberal husband. Whilst she is easily persuaded by her husband, she often disapproves of …
[16]	he would die in captivity. Among her tutors during her youth was her mother's sister, Maria Dolores Molina, who was the mother of her first cousin and future husband Manuel Luis Quezon. After her father's imprisonment, she …
[17]	102nd birthday at her California residence. She died of natural causes at her home in Toluca Lake, California, on September 19, 2011. She had been in relatively good health until a few months before her death. […]
[18]	through the intervention of the Spanish ambassador in Berlin that they were released. After reaching Paris, Princess Dolores and her husband moved permanently to Spain. They settled in Seville where Princess Dolores gave …
[19]	that Briscoe would return. Servants reported that Mrs Briscoe would go out riding where she would meet Gordon, but Gordon would see her home but never enter the house with her. Gordon eloped with Mrs Briscoe on 21 October …
[20]	age of 100 at his home in Toluca Lake, California. His grandson Zach Hope told TV interviewer Soledad O'Brien that, when asked on his deathbed where he wanted to be buried, Hope told his wife, Dolores, "Surprise me." […]

Reranker	Selected documents	Answer	Reranker	Selected documents	Answer
Initial Retrieval	[1] [2] [3] [4] [5]	Inside a well	RankZephyr (7B)	[2] [20] [17] [14] [9]	Home
BGE-Reranker$_{L.}$	[17] [2] [20] [1] [8]	Home	Rearank (7B)	[20] [17] [2] [1] [9]	Home
MonoT5 (3B)	[1] [2] [17] [20] [13]	In a well	ReasonRank (7B)	[17] [20] [2] [1] [8]	Home
RankT5 (3B)	[17] [20] [1] [2] [13]	Home	Rank4Gen (8B)	[20] [2] [17]	Home
RankLlama (7B)	[2] [17] [20] [1] [3]	Home	SetR (8B)	[2] [17] [20]	Home
RankVicuna (7B)	[2] [1] [11] [3] [14]	A well	RubricRanker (8B)	[20] [2] [17]	Home
AdaTutoRank (8B)	[17] [20]	**Toluca Lake**			

Table 11: Case study on 2WikiMultihopQA for the two-hop query *Where did Dolores Hope's husband die?* The upper block lists the 20 BM25 candidates, where blue marks the spans that establish the answer: passage 20 states that he died at his home in Toluca Lake, and passage 17 gives the same address as the couple's home. The lower block lists the documents each method feeds to the generator (the top-5 for a ranking method, the chosen subset for a set selector), where green marks a selected passage that supports the answer and red one that does not. […] marks an elision inside a passage.

Benchmark: PopQA **Query:** What genre is Whitney? **Gold answer:** *pop music*

ID	Candidate document
[1]	Whitney Awards The Whitney Awards are awards given annually for novels by LDS authors. Established in 2007, they are named after Orson F. Whitney, a prominent early member of the LDS Church. There are several categories for …
[2]	James Whitney (filmmaker) "For other people named James Whitney, see James Whitney (disambiguation)" James Whitney (December 27, 1921 – April 8, 1982), younger brother of John, was a filmmaker regarded as one of the great …
[3]	singing Vietnamese songs. Uyen Linh's idol is Whitney Houston and her favorite genre is RnB. According to one of her interviews, she admires the voice of Whitney because "Whenever Linh hears her singing, Linh feels like …
[4]	her book "The Mystery of the Haunted Pool" won an Edgar Award from the Mystery Writers of America for Best Juvenile novel, and she duplicated the honor in 1964, for "The Mystery of the Hidden Hand". In 1988, the MWA gave her …
[5]	romance novel genre, create a community of female readers sharing in similar emotions and knowledge. After the initial publication in 1985, "Whitney, My Love" was adapted by a number of publishers to at least 6 different …
[6]	Phyllis A. Whitney Phyllis Ayame Whitney (September 9, 1903 – February 8, 2008) was a Japanese-born American mystery writer. Rare for her genre, she wrote mysteries for both the juvenile and the adult markets, many of which …
[7]	sway over so many genres. But without a gospel single from "The Preacher's Wife" soundtrack — some of her most emotive work — this isn't Whitney at her best." AllMusic reviewer Stephen Thomas Erlewine felt that "The Essential Whitney Houston" plays much like "The Greatest Hits" […]
[8]	record vocals on an anti-bullying compilation album entitled "All About Bullies". The Album went on to win a Grammy Award at the 54th Annual Grammy Awards for "Best Children's Album" as well as Parent's Choice Gold Award in …
[9]	genre or style, was in very few permanent collections when she was alive. Since her death, Wilke's work has been acquired into the permanent collections of The Museum of Modern Art, New York, the Whitney Museum of American …
[10]	revised memorials and take expeditions through what was then known as Indian Territory to support his cause. Later Whitney's dream was realized through the efforts of Theodore Judah. In the end, Whitney lived to see his …
[11]	an engagement present. Later, a bus crashes into the market and Whitney is trapped underneath it. Whitney is rescued and in hospital, she talks to Mick about her marriage. He comforts her and she kisses him, which is seen by …
[12]	number 1 hit (US R&B) with his song "Sensitivity", with another on the "House Party 2" soundtrack "Yo Baby Yo". Also in 1990 pop singer Whitney Houston recorded "I'm Your Baby Tonight", produced by Babyface and his new jack swing producing partner […]
[13]	father. To the surprise of her small town, Whitney returns as a sophisticated young lady. Furthermore, possessing what she believes is a substantial inheritance from her grandmother, Whitney is determined to marry Paul. Once …
[14]	a grandson of George Macculloch Miller (1832–1917), the founder what would become the United Hospital Fund. The marriage to "Cully" Miller was long and happy, and Whitney had two more children: She died on July 18, 1986 at …
[15]	the "Tribune" until its demise, but Whitney and his advisors controlled the paper. Whitney initially left management of the newspaper to Walter Thayer, a longtime advisor. Thayer did not believe the "Tribune" was a financial …
[16]	the next day, and Petersmeyer went on to become a partner at the firm and run Whitney Communications. Whitney had been investing since the 1930s, founding Pioneer Pictures in 1933 and acquiring a 15% interest in Technicolor …
[17]	tours of Japan and a full-length album release followed, also entitled "I Am What I Am". In 2007, 2008 and 2009, the tour was also brought to Europe where she maintained a cult following. In December 2009, Whitney had a …
[18]	They also manufactured milling machines and twist drills. What remains of the original Pratt & Whitney is now Pratt & Whitney Measurement Systems, located in Bloomfield, Connecticut. Pratt & Whitney Measurement Systems is an …
[19]	ends their affair. Devastated when Tony accepts Bianca's marriage proposal, Whitney locks herself in her bedroom. Not knowing what to do, Bianca accepts Dr Poppy Merritt's (Amy Darcy) help and she refers Whitney to a …
[20]	the institutions where art is displayed. In a review of Warhol's 1971 retrospective show at the Whitney, she observed that cows are a common subject of genre paintings that people display in their homes, and that the …

Reranker	Selected documents	Answer	Reranker	Selected documents	Answer
Initial Retrieval	[1] [2] [3] [4] [5]	Abstract Cinema	RankZephyr (7B)	[3] [5] [6] [4] [1]	Romantic Novels
BGE-Reranker$_{L.}$	[6] [2] [1] [5] [3]	Romantic Novels Of Suspense	Rearank (7B)	[3] [7] [12] [6] [5]	Mystery Writer Genre
MonoT5 (3B)	[1] [6] [5] [2] [3]	Mystery Writer Genre	ReasonRank (7B)	[3] [6] [5] [1] [2]	Romantic Novels
RankT5 (3B)	[6] [1] [4] [5] [2]	Romantic Suspense Novels	Rank4Gen (8B)	[7] [6] [12]	Mystery Writer
RankLlama (7B)	[1] [6] [4] [2] [5]	Romantic Novels	SetR (8B)	[1] [4] [6] [7] [12]	Romantic Novels
RankVicuna (7B)	[1] [5] [6] [4] [2]	Romantic Novels	RubricRanker (8B)	[6] [4]	Romantic Suspense Novels
AdaTutoRank (8B)	[7] [12]	**Pop**			

Table 12: Case study on PopQA for the query *What genre is Whitney?*, which asks about the Whitney Houston album. The upper block lists the 20 BM25 candidates, where blue marks the span that establishes the answer. The lower block lists the documents each method feeds to the generator (the top-5 for a ranking method, the chosen subset for a set selector), where green marks a selected passage about that Whitney and red one about a different entity of the same name; the pool holds at least seven such entities, including two novelists, a filmmaker, a museum and a publisher. […] marks an elision inside a passage.

Query: Which affiliate left the company that owns and operates most of the CBC television stations due to an agreement with CHBC?
Gold answer: *CFJC-TV in Kamloops*

Dim.	Query-specific rubrics
Rel.	Does each document directly discuss the departure of CFJC-TV in Kamloops from the company that owns and operates most of the CBC television stations, specifically focusing on the agreement with CHBC that caused this departure, rather than merely mentioning CBC affiliates, CHBC, or unrelated affiliation changes?
Aut.	(i) For each document that identifies the affiliate that left the company that owns and operates most CBC television stations, does it correctly state that it was CFJC-TV in Kamloops, and not some other station? (ii) For each document that describes the departure of CFJC-TV in Kamloops from that company, does it attribute the departure to an agreement with CHBC, rather than any other reason?
Qua.	Is each document clearly structured and focused enough that a reader can directly extract the fact that CFJC-TV in Kamloops is the affiliate that left the company owning and operating most CBC television stations due to an agreement with CHBC, rather than having this specific relationship buried in unrelated, disorganized, or superficial content?
Cmp.	Does the document set collectively provide both the identity of the company that owns and operates most CBC television stations AND the specific fact that CFJC-TV in Kamloops left that company due to an agreement with CHBC?
Red.	Does the document set avoid redundancy where multiple documents merely repeat the core fact that CFJC-TV in Kamloops left the company that owns and operates most of the CBC television stations due to an agreement with CHBC, without providing any incremental information or context?
Con.	Across the document set, is there no factual contradiction regarding the identity of the departing affiliate (consistently identifying CFJC-TV in Kamloops) and the reason for departure (an agreement with CHBC), with no conflicting information such as a different station or city being named for the same event?
Cpl.	Does the document set as a whole provide the identity of the company that owns and operates most CBC television stations, state that an affiliate left this company due to an agreement with CHBC, and explicitly identify that affiliate as CFJC-TV in Kamloops?
Den.	Across the documents in the set, does the text that specifically identifies CFJC-TV in Kamloops as the affiliate that left the company owning and operating most CBC television stations due to an agreement with CHBC occupy the majority of each document's length, rather than each document being largely filled with unrelated content (e.g., other CBC/CHBC history, or unrelated station details) where only a small fraction of the text is actually useful for this specific affiliate change?
Rea.	Using only the document set and no external knowledge, can the full reasoning chain be completed — identifying the company that owns and operates most CBC television stations, establishing that an affiliate left this company because of an agreement with CHBC, and concluding that this affiliate was CFJC-TV in Kamloops?

ID	Candidate document
[1]	[. . .] One private CBC affiliate, CHBC-TV in Kelowna, joined E! (then known as CH) on February 27, 2006. When a private CBC affiliate reaffiliated with another network, the CBC normally added a retransmitter of its nearest O&O station. However, due to an agreement between CHBC and CFJC-TV in Kamloops, CFJC also disaffiliated from the CBC on February 27, 2006, but no retransmitters were installed in the licence area.
[2]	[. . .] In late 2003, the CBC notified CHBC that it did not intend to renew its affiliation agreement with the station after it expired in August 2005. In response, the station filed an application with the Canadian Radio-television [. . .]
[3]	[. . .] and Telecommunications Commission in 2004 to disaffiliate from the CBC; the CRTC gave approval to the disaffiliation on February 28, 2005. [. . .] After its BCI-TV partner CFJC-TV received similar approval to disaffiliate from the CBC, both stations switched affiliations on February 27, 2006 and continued the operation of BCI-TV [. . .]
[4]	CICT, CHBC, CHEX, CISA and CKWS launched in the 1950s as CBC Television affiliates, while CHAN-TV launched in 1960 and soon became Vancouver's original CTV affiliate. All of these were eventually supplanted by . . .
[5]	had left CBC Television solely dependent on cable and satellite carriage of its Vancouver station CBUT in the market, with no new terrestrial transmitters being installed in the Kamloops area. The Canadian . . .
[6]	of Canada's Weatheradio Canada service. The CBC operates two national broadcast television networks; CBC Television in English, and Ici Radio-Canada Tele in French. Like private broadcasters, both those . . .
[7]	were produced live to air. Locally produced programs during the station's early days included "Kids Bids", "The Three R's", "Romper Room", "Let's Visit", "Midday", "Focus" and "Okanagan Magazine". In 1964, . . .
[8]	its other E! owned-and-operated stations, stating that "a second conventional TV network [was] no longer key to the long-term success" of the company. Although for a time it was reported that CHBC might cease . . .
[9]	via an aerial signal in the Hamilton-Toronto-Buffalo area on CHCH-DT Channel 18. CHCH retains this digital signal under its new ownership. In addition, CHBC, CHEK and CJNT have since converted to digital . . .
[10]	is also a high definition feed available on Telus TV channel 692, and Shaw Direct Classic tier channel 7 and Advanced tier channel 507. The station first signed on the air on April 8, 1957 as CFCR-TV, . . .
[11]	CHEX was originally a CBC affiliate, but after 60 years with the CBC it began airing CTV programming on August 31, 2015 under a programming supply agreement. On August 14, 2018, it was announced that CHEX's . . .
[12]	[. . .] Kelowna's CHBC and Kamloops's CFJC, the latter owned by the Jim Pattison Group, also disaffiliated from the CBC in February 2006 and joined CH. Although CFJC was not owned by Canwest, its joint sales agreement with CHBC necessitated its affiliation switch.
[13]	CKX-TV CKX-TV was a television station in Brandon, Manitoba, Canada, formerly affiliated with CBC Television. Owned and operated by CTVglobemedia, it was the first privately owned television station in . . .
[14]	Durham Region. The station remained affiliated with CBC despite the fact its signal overlaps with that of the network's Toronto owned-and-operated station (O&O) CBLT-DT; as a result, the Toronto market was . . .
[15]	ownership of the station changed, beginning with the purchase of CKOK's one-third ownership by general manager Roy Chapman, which he later sold to CHAN. Selkirk Communications brought CJIB, and along with it, . . .
[16]	[. . .] Most private affiliates produce their own local newscasts for a duration of at least 35 minutes. Some of the private affiliates later began adding CBC's overnight programming to their schedules since the network began broadcasting 24 hours a day in October 2006. Following the disaffiliation of the last privately-owned CBC affiliate CKSA-DT in Lloydminster on August 31, 2016, no more private stations operate as CBC affiliates [. . .]
[17]	of media consolidation have always been allowed. In small markets where the population could not adequately support multiple television stations competing for advertising revenue, the CRTC began permitting . . .
[18]	CHBC-DT CHBC-DT, virtual channel 2 (UHF digital channel 27), is a Global owned-and-operated television station located in Kelowna, British Columbia, Canada. The station is owned by Corus Entertainment. CHBC . . .
[19]	Telecommunications Commission (CRTC) has significantly more lenient rules regarding media ownership. As such, most television stations, regardless of market size, are now O&Os of their respective networks, . . .
[20]	compete with CBC. The deal was approved by the Competition Bureau in March 2013, and by the CRTC in June 2013. On October 28, 2015, the CRTC made public an application by Bell to disaffiliate CJDC from CBC . . .

Reranker	Selected documents	Rel.	Aut.	Qua.	Cmp.	Red.	Con.	Cpl.	Den.	Rea.	Doc.	Set	Glo.	Ovr.
RubricRanker (8B)	[1] [3] [12] [2]	2.5	2.5	2.0	1.0	2.0	4.0	4.0	2.0	4.0	2.33	2.33	3.33	2.67
AdaTutoRank (8B)	[1] [12]	**4.0**	**4.0**	**3.0**	**2.0**	**2.0**	**4.0**	**4.0**	**2.0**	**4.0**	**3.67**	**2.67**	**3.33**	**3.22**

Table 13: Setwise-level case study on SetwiseEvalKit: the query-specific rubrics, the top-20 BM25 candidates, and each selector's set with its judge scores (0–4).

F Details of Prompt

F.1 AdaTutoRanker Prompt: RAG Setting

System: You are a high-quality document set selector for a retrieval-augmented QA system. Given a user question and a list of retrieved passages, select the best subset that jointly answers the question, then rank the selected passages from most to least useful.
Use these 9 rubrics to judge a candidate document set:
Relevance (doc-level, primary objective): Does the document directly address, at the topic level, the core information need of the query (naming the concrete subject/facet the query asks about), rather than merely matching surface keywords?
Authenticity (doc-level): Are the document's factual statements consistent with the correct reference facts (dates, numbers, names, causal conclusions) given by the query/answer? Drop documents that state falsehoods.
Quality (doc-level): Is the document clearly structured, focused, and information-dense enough that the target information can be directly extracted, rather than being empty, too shallow, noisy, or low-quality filler?
Complementarity (set-level): Do the retained documents complement one another so that, taken together, they cover the key information elements the answer requires more completely than any single document could? Prefer subsets whose documents jointly raise coverage.
Redundancy (set-level): Does the subset avoid two or more documents that convey essentially the same core information with no incremental value? When documents duplicate each other, keep only the single best one.
Conflict (set-level): Are the documents free of mutual factual contradiction on the same point, so the subset consistently supports the correct answer? Reject subsets that contain contradictory statements.
Completeness (global-level): Does the whole subset collectively supply every key information element the correct answer depends on, leaving no obvious gap?
Density (global-level): Across the subset, does the query-relevant content occupy the majority of each document's length, rather than the documents being bloated with irrelevant padding around a small useful snippet?
Reachability (global-level): Using only this subset and no external knowledge, can the full reasoning chain required to produce the correct answer be completed — is every link grounded in the set?
Output format (STRICT): respond with only the selected document identifiers, each wrapped in square brackets and separated by a single space, ordered from most to least useful. Example: [4] [2] [7] [11] [1]. Do not output any other text or explanation.

User: I will provide you with **<NUM>** passages, each indicated by a numerical identifier []. Select and rank the passages based on their joint usefulness for the user question: **<QUESTION>**.
[PASSAGES]
<CONTEXT>
User Question: **<QUESTION>**.
Select the best subset of the **<NUM>** passages above so that, judged by the 9 dimensions rubrics, the selected set best answers the question. Then rank the selected passages from most to least useful. The number of selected passages is at most 10 and should not include redundant ones. The output must be only the selected identifiers in ranked order, e.g. [2] [1] [4]. Only respond with the selection; do not say any other word and do not explain.

F.2 AdaTutoRanker Prompt: Deep-Research Setting

System: You are a high-quality document set selector for a deep-research agent. Given a search query from the agent, the agent's reasoning as query intent, and a list of retrieved documents, select the best subset that jointly answers the query, then rank the selected documents from most to least useful.
The query intent may clarify why this query was issued and what information it is trying to find, but it may also contain broader, exploratory, or irrelevant thoughts. Stay anchored on the query itself and only use the parts of the query intent that are actually relevant to this query.
Use these 9 rubrics to judge a candidate document set:
Relevance (doc-level, primary objective): Does the document directly address, at the topic level, the core information need of the query (naming the concrete subject/facet the query asks about), rather than merely matching surface keywords?
Authenticity (doc-level): Are the document's factual statements consistent with the correct reference facts (dates, numbers, names, causal conclusions) given by the query/answer? Drop documents that state falsehoods.
Quality (doc-level): Is the document clearly structured, focused, and information-dense enough that the target information can be directly extracted, rather than being empty, too shallow, noisy, or low-quality filler?
Complementarity (set-level): Do the retained documents complement one another so that, taken together, they cover the key information elements the answer requires more completely than any single document could? Prefer subsets whose documents jointly raise coverage.
Redundancy (set-level): Does the subset avoid two or more documents that convey essentially the same core information with no incremental value? When documents duplicate each other, keep only the single best one.
Conflict (set-level): Are the documents free of mutual factual contradiction on the same point, so the subset consistently supports the correct answer? Reject subsets that contain contradictory statements.
Completeness (global-level): Does the whole subset collectively supply every key information element the correct answer depends on, leaving no obvious gap?
Density (global-level): Across the subset, does the query-relevant content occupy the majority of each document's length, rather than the documents being bloated with irrelevant padding around a small useful snippet?
Reachability (global-level): Using only this subset and no external knowledge, can the full reasoning chain required to produce the correct answer be completed — is every link grounded in the set?
Output format (STRICT): respond with only the selected document identifiers, each wrapped in square brackets and separated by a single space, ordered from most to least useful. Example: [4] [2] [7] [11] [1]. Do not output any other text or explanation.

User: I will provide you with **<NUM>** passages, each indicated by a numerical identifier []. Select and rank the passages based on their joint usefulness for the search query: **<QUERY>**.
[QUERY INTENT]
<QUERY_INTENT>
[PASSAGES]
<CONTEXT>
Search Query: **<QUERY>**.

Select the best subset of the <NUM> passages above so that, judged by the 9 dimensions rubrics, the selected set best answers the query (anchored on the query, with the query intent only as clarification). Then rank the selected passages from most to least useful. The number of selected passages is at most 10 and should not include redundant ones. The output must be only the selected identifiers in ranked order, e.g. [2] [1] [4]. Only respond with the selection; do not say any other word and do not explain.

F.3 Rubric Generation Prompt: Short-form Scenario

System: You are a senior rubric-generation expert. Given a (query, answer) pair, you generate binary (Yes/No) evaluation rubrics along nine dimensions. These rubrics are used to judge whether a candidate document set can effectively support answering the query and producing the correct answer.
The nine dimensions come in three families, and each family is scoped differently. Respect the scope of the family a dimension belongs to:
- Doc-level (Relevance, Authenticity, Quality): each rubric judges EACH individual document in the set. Phrase them as "Does each document . . . ".
- Set-level (Complementarity, Redundancy, Conflict): each rubric judges the relationships BETWEEN documents in the set. Phrase them as "Does the document set . . . " / "Across the documents, . . . ".
- Global-level (Completeness, Density, Reachability): each rubric judges the ENTIRE set as one unit — its overall utility for answering the query. Phrase them as "Does the whole document set . . . ".
Rubric requirements
Every rubric you generate MUST satisfy all of the following:
- Relevant: It must be directly tied to the core information need of the query and help produce the answer, so that it meaningfully reduces uncertainty when judging quality.
- Binary: It must be a Yes/No question. No degree-based phrasing (e.g., "to what extent").
- Qualitative & query-specific: It must be purpose-built for this specific (query, answer). You MUST name the concrete entities, facts, dates, or numbers from the query/answer. Never use vague wording such as "relevant content" or "important information".
- Grounded: Its judgment criterion must be derivable from the given (query, answer); a judge should not need external knowledge to answer it.
- Salient: It should focus on the aspects an experienced information-retrieval expert would care about.
- Correctly scoped: It must match the scope (doc-level / set-level / global-level) of its own dimension, as defined above.
- Weighted: It must carry an integer weight from 1 to 5 stating how important this rubric is for judging whether the document set can support answering THIS query.
Weight scale
Assign the weight by asking: if the document set fails this rubric, how badly is the query damaged?
- 1 = marginally useful; minor added value only
- 2 = useful but secondary
- 3 = moderately important; clearly helpful for evaluating a good document set
- 4 = very important; strong effect on document set quality
- 5 = critical; if missing or violated, the document set is seriously inadequate for answering the query
Weight the rubric, not the dimension: two rubrics under the same dimension can differ in weight, and a rubric under a "minor" dimension can outweigh one under a "major" dimension if this particular query depends on it more. Spread the weights — do not label everything 4 or 5. Reserve 5 for the few rubrics that are genuinely make-or-break for this query, and use 1-2 for the nice-to-have ones.
Dimensions
Doc-level
- Relevance: Measures how semantically related each document's content is to the core information need of the query — whether the document directly addresses, at the topic level, what the query is asking, rather than merely matching surface keywords.
Requirement: The rubric must name the concrete subject and information facet the query asks about (e.g., specific people / works / events / time dimension) and judge whether the document truly discusses this specific question, not merely whether it is "relevant".
Goal: Keep documents that directly answer the user's question; discard off-topic or keyword-only, semantically irrelevant documents.
- Authenticity: Measures how well the facts stated in each document conform to objective real-world information.
Requirement: The rubric must spell out the correct reference fact (taken from the answer or the deterministic information in the query) and judge whether the document's statements are consistent with it, rather than asking whether the document "contains correct information". Only generate this dimension when the query/answer involves concrete, verifiable facts (e.g., dates, numbers, names, causal conclusions); otherwise return an empty list for it.
Goal: Discard documents with false statements or factual errors.
- Quality: Measures how usable each document is in terms of information organization and presentation — whether it is clearly structured, focused, and lets the reader directly extract the target information, rather than being poorly formatted, too short / shallow, noisy, or low-quality filler.
Requirement: The rubric must reference the specific information a reader needs to extract in order to answer this query (name the concrete entities, not generic terms), and judge whether the document makes that information clearly accessible. This dimension evaluates presentation quality only, not factual correctness.
Goal: Discard low-quality documents (empty content, low information density, unclear phrasing, etc.).
Set-level
- Complementarity: Measures the degree to which the documents in the set complement one another — whether, taken together, they cover all the key information elements the answer requires and jointly build a more complete, well-rounded answer than any single document could (e.g., one document gives an overview while another gives the details; one covers 70% of the must-have points while another covers the remaining 30%).
Requirement: The rubric must name the concrete key information elements the query/answer requires, and judge whether the set as a whole covers them across its documents. Distinguish complementarity (documents combine to increase information value) from conflict (documents contradict each other).
Goal: Reward sets whose documents jointly raise coverage of the query's information need when no single document fully satisfies it.
- Redundancy: Measures the degree of content redundancy among the documents in the set — whether two or more documents convey essentially the same core information, data, or conclusions without any incremental contribution (common when documents merely reprint, summarize, rewrite, or re-report the same source).
Requirement: The rubric must name the concrete query-specific information, and judge whether the set avoids having two or more documents that carry essentially identical information with no incremental value (in which case only the single best — highest information density, most concise, clearest — should be kept).
Goal: Discourage sets that pad in duplicated documents instead of adding new information.

- Conflict: Measures the degree of factual consistency across the documents in the set — whether different documents give mutually contradictory factual statements about the same point (e.g., two different values for the same fact, one correct and one wrong).
Requirement: The rubric must name the concrete fact at issue and judge whether the documents' statements are free of mutual contradiction, so that the set consistently supports the correct answer. A high-quality set must never contain factual contradictions. Only generate this dimension when the query/answer involves concrete, verifiable facts that documents could contradict; otherwise return an empty list for it.
Goal: Reject sets containing factual contradictions or documents with erroneous information.
Global-level
- Completeness: Measures whether the document set, as a whole, can COMPLETELY answer the query with no obvious information gaps — every key information element the correct answer depends on is present somewhere in the set.
Requirement: The rubric must enumerate the concrete key information elements the query/answer requires and judge whether the whole set collectively supplies all of them, leaving no obvious gap that would prevent producing the correct answer.
Goal: Reward sets that fully cover the query's information need; reject sets that miss essential pieces.
- Density: Measures, in terms of LENGTH / span, how much of each document is actually useful for answering this query — i.e., the ratio of useful content to the document's total length. It is NOT about whether a document is removable, and NOT about whether the set as a whole is complete. It focuses on the internal "signal-to-length" ratio of the documents.
- What we WANT (high Density): each document in the set is compact and on-point, so that the text carrying the query-relevant information (name the concrete query-specific facts/entities/dates) occupies the MAJORITY of that document's length.
- What we DO NOT want (low Density): a document that is long but mostly filler — the bulk of its length is off-topic / unrelated content, and only a small fraction of the text actually contributes useful information for this query.
Requirement: The rubric must name the concrete query-specific useful information, and judge whether — across the documents in the set — the useful content takes up the main portion of each document's length, rather than the documents being padded with large amounts of irrelevant text where only a small snippet is useful.
Goal: Reward sets whose documents are dense with useful content; penalize sets containing bloated documents that are mostly irrelevant padding with only a sliver of useful information.
- Reachability: Measures whether a model WITHOUT any external knowledge could complete the full reasoning chain required to produce the correct answer using ONLY the document set. Every link of the inference chain must be grounded in the set.
Requirement: The rubric must name the concrete reasoning steps / intermediate facts the query requires (e.g., the bridge entities and the final comparison), and judge whether the whole set self-containedly provides every link so the correct answer is derivable without outside knowledge.
Goal: Reward sets that are self-contained for the full reasoning chain; reject sets with a missing link that forces reliance on external knowledge.
Additional criteria
- Knowledge-cutoff: Do not ask for or rely on information that would violate a knowledge cutoff date.
- Coverage & non-redundancy: The rubrics together should cover the important aspects of the query without redundant or duplicate questions. In particular, do not restate the same question under two different dimensions — each dimension must contribute a distinct judgment.
- Generate at most 3 questions per dimension (fewer is fine). Every question must be important, useful, and non-redundant.
Example
Query: Which was founded earlier, Arthur's Magazine or First for Women?
Answer: Arthur's Magazine

```
{
  "Relevance": [
    {"rubric": "Does each document directly discuss the founding/launch date or early publication history of Arthur's Magazine or First for Women, rather than merely containing words like 'Arthur', 'magazine', or 'women' without addressing founding dates?", "weight": 5}
  ],
  "Authenticity": [
    {"rubric": "For each document, are its founding-date statements consistent with objective fact - namely Arthur's Magazine = 1844 and First for Women = 1989 - without misstating the years or attributing them to the wrong magazine?", "weight": 5},
    {"rubric": "For each document that states which was founded first, is that conclusion consistent with the fact that Arthur's Magazine predates First for Women, rather than claiming First for Women is older?", "weight": 3}
  ],
  "Quality": [
    {"rubric": "Is each document clearly structured and focused enough that a reader can directly extract the founding year of Arthur's Magazine and/or First for Women (or the conclusion that Arthur's Magazine came first), rather than having these key dates buried under large amounts of irrelevant content?", "weight": 2}
  ],
  "Complementarity": [
    {"rubric": "Does the document set contain at least one document that explicitly records the founding year of Arthur's Magazine AND at least one that explicitly records the founding year of First for Women, so that the two can be compared to determine which was founded first?", "weight": 5}
  ],
  "Redundancy": [
    {"rubric": "Does the document set avoid having two or more documents that convey essentially the same founding-date information (e.g., all merely stating 'Arthur's Magazine was founded in 1844') without any incremental information contribution?", "weight": 2}
  ],
  "Conflict": [
    {"rubric": "Are the statements across the document set about the two magazines' founding dates free of mutual contradiction (e.g., one saying 1844 and another saying 1850), so that they consistently support the conclusion that Arthur's Magazine was founded earlier?", "weight": 4}
  ],
  "Completeness": [
    {"rubric": "Does the document set as a whole provide the founding year of BOTH Arthur's Magazine and First for Women, leaving no gap that would prevent determining which was founded earlier?", "weight": 5}
  ],
  "Density": [
    {"rubric": "Across the documents in the set, does the text that actually states the founding date(s) of Arthur's Magazine and First for Women make up the majority of each document's length, rather than each document
```

```
being dominated by unrelated content (e.g., editorial history, staff, or coverage of other magazines) with only
a small fraction of its length mentioning those founding dates?", "weight": 2}
  ],
  "Reachability": [
    {"rubric": "Using only the document set and no external knowledge, can the full reasoning chain be completed
--- obtaining the founding year of Arthur's Magazine, obtaining the founding year of First for Women, and
comparing them --- to reach the conclusion that Arthur's Magazine was founded earlier?", "weight": 4}
  ]
}
```

Output format

Output ONLY a valid JSON object (no markdown code fences, no extra commentary) with exactly these nine keys: "Relevance", "Authenticity", "Quality", "Complementarity", "Redundancy", "Conflict", "Completeness", "Density", "Reachability". Each key maps to a list of objects, each object having exactly two fields: "rubric" (the Yes/No question string) and "weight" (an integer from 1 to 5). A dimension may have multiple rubrics. If a dimension does not apply (e.g., Authenticity or Conflict for a query with no verifiable facts), use an empty list [] — but still include the key.

User: Query: **<QUESTION>**
Answer: **<ANSWER>**
Rubrics (JSON):

F.4 Rubric Generation Prompt: Long-form Scenario

System: You are a senior rubric-generation expert. Given a (query, answer) pair, you generate binary (Yes/No) evaluation rubrics along nine dimensions. These rubrics are used to judge whether a candidate document set can effectively support answering the query and producing the correct answer.

The nine dimensions come in three families, and each family is scoped differently. Respect the scope of the family a dimension belongs to:
- Doc-level (Relevance, Authenticity, Quality): each rubric judges EACH individual document in the set. Phrase them as "Does each document . . . ".
- Set-level (Complementarity, Redundancy, Conflict): each rubric judges the relationships BETWEEN documents in the set. Phrase them as "Does the document set . . . " / "Across the documents, . . . ".
- Global-level (Completeness, Density, Reachability): each rubric judges the ENTIRE set as one unit — its overall utility for answering the query. Phrase them as "Does the whole document set . . . ".

Rubric requirements

Every rubric you generate MUST satisfy all of the following:
- Relevant: It must be directly tied to the core information need of the query and help produce the answer, so that it meaningfully reduces uncertainty when judging quality.
- Binary: It must be a Yes/No question. No degree-based phrasing (e.g., "to what extent").
- Qualitative & query-specific: It must be purpose-built for this specific (query, answer). You MUST name the concrete entities, facts, dates, or numbers from the query/answer. Never use vague wording such as "relevant content" or "important information".
- Grounded: Its judgment criterion must be derivable from the given (query, answer); a judge should not need external knowledge to answer it.
- Salient: It should focus on the aspects an experienced information-retrieval expert would care about.
- Correctly scoped: It must match the scope (doc-level / set-level / global-level) of its own dimension, as defined above.
- Weighted: It must carry an integer weight from 1 to 5 stating how important this rubric is for judging whether the document set can support answering THIS query.

Weight scale

Assign the weight by asking: if the document set fails this rubric, how badly is the query damaged?
- 1 = marginally useful; minor added value only
- 2 = useful but secondary
- 3 = moderately important; clearly helpful for evaluating a good document set
- 4 = very important; strong effect on document set quality
- 5 = critical; if missing or violated, the document set is seriously inadequate for answering the query

Weight the rubric, not the dimension: two rubrics under the same dimension can differ in weight, and a rubric under a "minor" dimension can outweigh one under a "major" dimension if this particular query depends on it more. Spread the weights — do not label everything 4 or 5. Reserve 5 for the few rubrics that are genuinely make-or-break for this query, and use 1-2 for the nice-to-have ones.

Dimensions

Doc-level

- Relevance: Measures how semantically related each document's content is to the core information need of the query — whether the document directly addresses, at the topic level, what the query is asking, rather than merely matching surface keywords.
Requirement: The rubric must name the concrete subject and information facet the query asks about (e.g., specific people / works / events / time dimension) and judge whether the document truly discusses this specific question, not merely whether it is "relevant".
Goal: Keep documents that directly answer the user's question; discard off-topic or keyword-only, semantically irrelevant documents.
- Authenticity: Measures how well the facts stated in each document conform to objective real-world information.
Requirement: The rubric must spell out the correct reference fact (taken from the answer or the deterministic information in the query) and judge whether the document's statements are consistent with it, rather than asking whether the document "contains correct information". Only generate this dimension when the query/answer involves concrete, verifiable facts (e.g., dates, numbers, names, causal conclusions); otherwise return an empty list for it.
Goal: Discard documents with false statements or factual errors.
- Quality: Measures how usable each document is in terms of information organization and presentation — whether it is clearly structured, focused, and lets the reader directly extract the target information, rather than being poorly formatted, too short / shallow, noisy, or low-quality filler.
Requirement: The rubric must reference the specific information a reader needs to extract in order to answer this query (name the concrete entities, not generic terms), and judge whether the document makes that information clearly accessible. This dimension evaluates presentation quality only, not factual correctness.
Goal: Discard low-quality documents (empty content, low information density, unclear phrasing, etc.).

Set-level

- Complementarity: Measures the degree to which the documents in the set complement one another — whether, taken together, they cover all the key information elements the answer requires and jointly build a more complete, well-rounded answer than any single document could (e.g., one document gives an overview while another gives the details; one covers 70% of the must-have points while another covers the remaining 30%).
Requirement: The rubric must name the concrete key information elements the query/answer requires, and judge whether the set as a whole covers them across its documents. Distinguish complementarity (documents combine to increase information value) from conflict (documents contradict each other).
Goal: Reward sets whose documents jointly raise coverage of the query's information need when no single document fully satisfies it.
- Redundancy: Measures the degree of content redundancy among the documents in the set — whether two or more documents convey essentially the same core information, data, or conclusions without any incremental contribution (common when documents merely reprint, summarize, rewrite, or re-report the same source).
Requirement: The rubric must name the concrete query-specific information, and judge whether the set avoids having two or more documents that carry essentially identical information with no incremental value (in which case only the single best — highest information density, most concise, clearest — should be kept).
Goal: Discourage sets that pad in duplicated documents instead of adding new information.
- Conflict: Measures the degree of factual consistency across the documents in the set — whether different documents give mutually contradictory factual statements about the same point (e.g., two different values for the same fact, one correct and one wrong).
Requirement: The rubric must name the concrete fact at issue and judge whether the documents' statements are free of mutual contradiction, so that the set consistently supports the correct answer. A high-quality set must never contain factual contradictions. Only generate this dimension when the query/answer involves concrete, verifiable facts that documents could contradict; otherwise return an empty list for it.
Goal: Reject sets containing factual contradictions or documents with erroneous information.
Global-level
- Completeness: Measures whether the document set, as a whole, can COMPLETELY answer the query with no obvious information gaps — every key information element the correct answer depends on is present somewhere in the set.
Requirement: The rubric must enumerate the concrete key information elements the query/answer requires and judge whether the whole set collectively supplies all of them, leaving no obvious gap that would prevent producing the correct answer.
Goal: Reward sets that fully cover the query's information need; reject sets that miss essential pieces.
- Density: Measures, in terms of LENGTH / span, how much of each document is actually useful for answering this query — i.e., the ratio of useful content to the document's total length. It is NOT about whether a document is removable, and NOT about whether the set as a whole is complete. It focuses on the internal "signal-to-length" ratio of the documents.
- What we WANT (high Density): each document in the set is compact and on-point, so that the text carrying the query-relevant information (name the concrete query-specific facts/entities/dates) occupies the MAJORITY of that document's length.
- What we DO NOT want (low Density): a document that is long but mostly filler — the bulk of its length is off-topic / unrelated content, and only a small fraction of the text actually contributes useful information for this query.
Requirement: The rubric must name the concrete query-specific useful information, and judge whether — across the documents in the set — the useful content takes up the main portion of each document's length, rather than the documents being padded with large amounts of irrelevant text where only a small snippet is useful.
Goal: Reward sets whose documents are dense with useful content; penalize sets containing bloated documents that are mostly irrelevant padding with only a sliver of useful information.
- Reachability: Measures whether a model WITHOUT any external knowledge could complete the full reasoning chain required to produce the correct answer using ONLY the document set. Every link of the inference chain must be grounded in the set.
Requirement: The rubric must name the concrete reasoning steps / intermediate facts the query requires (e.g., the bridge entities and the final comparison), and judge whether the whole set self-containedly provides every link so the correct answer is derivable without outside knowledge.
Goal: Reward sets that are self-contained for the full reasoning chain; reject sets with a missing link that forces reliance on external knowledge.
Additional criteria
- Knowledge-cutoff: Do not ask for or rely on information that would violate a knowledge cutoff date.
- Coverage & non-redundancy: The rubrics together should cover the important aspects of the query without redundant or duplicate questions. In particular, do not restate the same question under two different dimensions — each dimension must contribute a distinct judgment.
- Generate at most 3 questions per dimension (fewer is fine). Every question must be important, useful, and non-redundant.
Example
Query: How does strain-softening behavior in sensitive clays differ from that in non-sensitive clays in terms of pore pressure development, shear strength, and strain effects?
Answer: In sensitive (e.g., quick) clays, undrained shearing produces POSITIVE excess pore pressures due to their contractive, collapsible structure, driving rapid strength loss through pore-pressure-induced softening, with strain tending to localize into shear bands. In contrast, non-sensitive / overconsolidated clays may develop NEGATIVE excess pore pressures (dilative tendency), show a more gradual shear-strength change, and deform more diffusely.


```
{
  "Relevance": [
    {"rubric": "Does each document directly address the strain-softening behavior of clays in terms of pore pressure development, shear-strength degradation, or strain localization (for sensitive and/or non-sensitive clays), rather than merely mentioning 'clay' or general soil mechanics without discussing strain-softening?", "weight": 5}
  ],
  "Authenticity": [
    {"rubric": "For each document, are its stated mechanisms consistent with established soil mechanics - namely that sensitive/quick clays generate positive excess pore pressure (contractive) under undrained shear while overconsolidated non-sensitive clays can develop negative excess pore pressure (dilative) - rather than asserting the opposite relationship?", "weight": 5},
    {"rubric": "For each document that describes strain effects, is its account consistent with strain localizing into shear bands in sensitive clays versus more diffuse deformation in non-sensitive clays, rather than reversing the two?", "weight": 3}
  ],
  "Quality": [
    {"rubric": "Is each document clearly structured and focused enough that a reader can directly extract the comparison of pore-pressure behavior, shear-strength change, and strain effects between sensitive and non-sensitive clays, rather than having these buried under large amounts of unrelated content?", "weight": 2}
  ],
  "Complementarity": [
```

```
    {"rubric": "Does the document set, taken together, cover all three required aspects - pore pressure
development, shear-strength degradation, and strain/localization effects - for BOTH sensitive and non-sensitive
clays (e.g., different documents supplying different aspects), so that the documents jointly build the full
comparison rather than any single aspect being missing across the set?", "weight": 5}
  ],
  "Redundancy": [
    {"rubric": "Does the document set avoid having two or more documents that convey essentially the same
finding (e.g., several documents merely restating that sensitive clays develop positive excess pore pressure
under undrained shear) without adding any incremental aspect such as shear-strength behavior, strain localization,
or the non-sensitive-clay contrast?", "weight": 2}
  ],
  "Conflict": [
    {"rubric": "Are the documents in the set free of mutual contradiction on the key mechanisms - e.g., not
one stating that sensitive clays develop positive excess pore pressure under undrained shear while another
asserts negative excess pore pressure for the same condition - so that the set consistently supports the correct
sensitive-vs-non-sensitive comparison?", "weight": 4}
  ],
  "Completeness": [
    {"rubric": "Does the document set as a whole cover every key information element the answer requires -
pore pressure development, shear-strength degradation, and strain/localization effects for BOTH sensitive and
non-sensitive clays - leaving no obvious gap that would prevent producing the full sensitive-vs-non-sensitive
comparison?", "weight": 5}
  ],
  "Density": [
    {"rubric": "Across the documents in the set, does the text that actually describes the strain-softening
mechanisms (pore pressure, shear strength, and strain effects for sensitive vs non-sensitive clays) make up the
majority of each document's length, rather than each document being dominated by unrelated content with only a
small fraction of its length addressing these mechanisms?", "weight": 2}
  ],
  "Reachability": [
    {"rubric": "Using only the document set and no external knowledge, can the full reasoning chain be completed
--- establishing that sensitive clays develop positive excess pore pressure and rapid, localized strength loss,
establishing the contrasting behavior of non-sensitive/overconsolidated clays (negative excess pore pressure,
gradual strength change, diffuse deformation), and contrasting the two --- to reach the correct differentiation?",
"weight": 4}
  ]
}
# Output format
Output ONLY a valid JSON object (no markdown code fences, no extra commentary) with exactly these nine keys: "Relevance",
"Authenticity", "Quality", "Complementarity", "Redundancy", "Conflict", "Completeness", "Density", "Reachability". Each key maps
to a list of objects, each object having exactly two fields: "rubric" (the Yes/No question string) and "weight" (an integer from 1 to 5). A
dimension may have multiple rubrics. If a dimension does not apply (e.g., Authenticity or Conflict for a query with no verifiable facts), use
an empty list [] — but still include the key.

User: Query: <QUESTION>
Answer: <ANSWER>
Rubrics (JSON):
```

F.5 Rubric-based Judge Scoring Prompt: Short-form Scenario

System: You are a senior search-result quality assessor. Given a user query, a candidate document set (doc set, variable size) and a list of rubrics (each labeled with its dimension type and judgment content), you must score every rubric on a 0–10 scale, strictly grounded in the document content.
This round evaluates all 9 dimensions at once, grouped into three categories:
[A. Doc-Level dimensions] — judged for EACH individual document in the doc set: Relevance, Authenticity, Quality.
[B. Set-Level dimensions] — judged for the ENTIRE doc set as a whole; give ONE overall score, not per-document: Complementarity, Redundancy, Conflict.
[C. Global-Level dimensions] — also judged for the ENTIRE doc set as a whole; give ONE overall score, not per-document: Completeness, Density, Reachability.
Absolute prerequisite: you MUST read the documents first
- Every score must be grounded in the actual document text. Judge only based on what ACTUALLY appears in the documents; do not use your own external knowledge to fill in information the documents do not state.
- The rubric text is itself query-specific (it already contains the concrete entities, facts, and numbers of the question); the user query is also provided above for context. Judge directly against the rubric, grounded in the documents.
Evaluation procedure — RELEVANCE GATE (read carefully, follow the order)
Think step by step in this order:
1. FIRST evaluate the Relevance dimension for EVERY document in the doc set (give each document its per_doc Relevance score).
2. Define the RELEVANT SUBSET = the documents whose Relevance score ¿= 1 (i.e. documents that actually touch the query's key information point(s)). Documents with Relevance = 0 are OFF-TOPIC and are EXCLUDED from EVERY other dimension: they NEVER count as providing coverage, as a distinct information point, as useful length, or as a comparable statement, and they NEVER count toward "no redundancy" or "no conflict".
3. Then check the relevance outcome:
   - CASE A — the RELEVANT SUBSET is EMPTY (every document's Relevance score = 0): the documents this ranker selected are all useless for this question, so the other 8 dimensions (Authenticity, Quality, Complementarity, Redundancy, Conflict, Completeness, Density, Reachability) carry NO evaluation meaning. In this case DO NOT score them — output each of those 8 dimensions as null (see output format). Only the Relevance dimension is scored.

- CASE B — the RELEVANT SUBSET is non-empty (at least ONE document has Relevance ¿= 1): evaluate the other 8 dimensions ONLY over the RELEVANT SUBSET, and apply each dimension's own APPLICABILITY PREREQUISITE below. A relevance-grounded dimension whose prerequisite is not met over the relevant subset (e.g. it needs a comparison / division of labour among relevant documents but there are not enough relevant documents, or no relevant document actually carries the key-information-point content it measures) MUST be output as null — NOT as a high score. Never reward a dimension merely because off-topic documents trivially "don't conflict / don't overlap / don't dilute".
4. Never leave Relevance itself null — Relevance is always scored.
Scoring scale (0–10)
Overall meaning (higher = better satisfies the rubric). Use the following anchors, and you MAY use intermediate integers (0–10) for in-between cases:
- 10 = Completely: fully satisfied, with sufficient detail and evidence.
- 8 = Mostly: largely satisfied, but missing some detail or support.
- 5 = Moderately: relevant content is mentioned, but key details are missing.
- 3 = Barely: not explicitly stated, only weakly inferable.
- 0 = Not at all: not addressed at all / irrelevant to the rubric / contradicts the facts.
I. Doc-Level dimensions: score EACH document individually
- Relevance: to what extent the document DIRECTLY discusses the specific information facet the rubric points to.
  - 10: the document's core content directly and fully discusses this specific question and can directly support the answer.
  - 8: directly discusses the topic, but the information is incomplete or unfocused.
  - 5: mentions the related topic only in passing, missing key information.
  - 3: barely discusses it directly, relevance is only weakly inferable.
  - 0: entirely off-topic, merely matching keywords or unrelated to the question.
- Authenticity: how well the facts stated in the document CONFORM to the reference fact given by the rubric. Only assess whether the STATEMENTS that appear are accurate; independent of relevance.
  - 10: the document explicitly states the relevant fact, fully consistent with the reference fact.
  - 8: the statement is consistent with the fact, but phrased unclearly or somewhat vaguely.
  - 5: partially correct, or the fact is right but wrapped in easily-misleading phrasing.
  - 3: only faintly touched upon; correctness is hard to judge or slightly off.
  - 0: the statement contradicts the reference fact, has an obvious factual error, or the document makes no statement related to that fact at all.
- Quality: how easy the document makes it for a reader to EXTRACT the target information, in terms of information organization and presentation. Assess presentation quality only, not factual correctness.
  - 10: clearly structured and focused; the target information is immediately apparent.
  - 8: mostly clear; the target information is extractable but mixed with some irrelevant content.
  - 5: the target information exists but is diluted by lots of irrelevant content and takes effort to find.
  - 3: messy layout, high noise; the target information is very hard to extract.
  - 0: empty, extremely noisy, or nearly unparseable content, OR the document is entirely off-topic and contains no target information at all.
II. Set-Level dimensions: give ONE overall score for the WHOLE doc set
- Complementarity: whether the key information points are covered through DIVISION OF LABOUR across DIFFERENT documents — i.e., different documents each contribute a DIFFERENT needed piece (one gives an overview, another the details; one supplies point P1, another supplies point P2), so that the whole is more valuable than any single document. APPLICABILITY PREREQUISITE (judge this FIRST): Complementarity is judged ONLY over the RELEVANT SUBSET (Relevance ¿= 1) and ONLY on the query-required key information points; off-topic documents (Relevance = 0) NEVER count as a contributing part. If FEWER THAN TWO relevant documents each contribute a DIFFERENT required point (there is at most one relevant document, or all relevant documents only carry the SAME single point), there is no possible cross-document division of labour on the key information points — Complementarity is NOT APPLICABLE: output it as null (NOT a high score). IMPORTANT: this dimension judges only the complementary STRUCTURE among the RELEVANT documents; it does NOT judge whether the set is complete. Do NOT lower the score merely because some required point is missing (that is Completeness's job) — an incomplete set can still be highly complementary if the points it DOES cover come from different documents. Conversely, if a SINGLE document already covers everything on its own, complementarity is LOW even though coverage may be perfect (there is no division of labour).
  - 10: strong complementarity — multiple documents clearly divide the labour, each contributing a DIFFERENT needed point, so the whole is far more valuable than any single document.
  - 8: mostly complementary — most of the covered points are supplied by different documents; only an occasional point is carried alone by one document.
  - 5: partial complementarity — there is some division of labour, but a large share of the covered content is concentrated in a single document.
  - 3: very weak complementarity — almost all useful content comes from one document, or several documents merely restate the same single point.
  - 0: no complementarity at all — only one document contributes any useful content (the others add nothing), OR a single document already covers everything by itself so no cross-document complementarity exists.
- Redundancy: judged against the MINIMAL document subset needed to cover the key information points the answer requires (the SAME key information points as Completeness). It checks whether each document brings INCREMENTAL coverage of those required information points, or merely repeats a point already covered by another document. If two or more documents cover the SAME required information point while adding no new required point, that overlap is redundancy — those documents could be dropped from the minimal covering set without losing any required point. (Example: if answering the query needs 2 information points, and both doc [1] and doc [5] only answer the SAME single point, that is redundancy.) Higher score = less redundancy (each document adds a new required point; close to a minimal covering set); lower = many documents duplicating the same point with no increment. APPLICABILITY PREREQUISITE (judge this FIRST): Redundancy is judged ONLY over the RELEVANT SUBSET (Relevance ¿= 1), against the required key information points; off-topic documents (Relevance = 0) are IGNORED and NEVER count as "a distinct point" nor improve the score. If FEWER THAN TWO relevant documents actually cover a key information point (there is nothing that could possibly overlap), Redundancy is NOT APPLICABLE: output it as null — do NOT output 10, and never reward "nothing to overlap".
  - 10: no redundancy — every document contributes a DISTINCT required information point; the set is essentially a minimal covering set with no duplicated point.
  - 8: nearly no redundancy — at most a negligible overlap; almost every document still adds a new required point.
  - 5: some redundancy — a few documents repeat an already-covered information point and add no new required point.
  - 3: heavy redundancy — many documents cover the same information point(s), giving little incremental coverage of the required points.
  - 0: severe redundancy — most documents duplicate the same one/few information points and bring no new required point (e.g., several documents all only answer the same single point).
- Conflict: whether different documents in the set are FREE OF mutual contradiction on the same factual point. APPLICABILITY PREREQUISITE (judge this FIRST): Conflict is judged ONLY over the RELEVANT SUBSET (Relevance ¿= 1); off-topic documents (Relevance = 0) never count as a comparable statement. Conflict can only be scored when the RELEVANT SUBSET contains AT LEAST

TWO comparable statements about the SAME query-relevant key-information-point factual point (i.e., two or more documents each make a checkable claim about the same fact, so that agreement or contradiction is even possible). If FEWER THAN TWO comparable statements exist about the same fact (the relevant fact is asserted by at most one document, or is not asserted at all), then there is nothing that could agree or contradict — Conflict is NOT APPLICABLE. In that case DO NOT give a 0–10 score; output this dimension as null (see output format). Do NOT give a high score merely because "nothing contradicts": absence of comparable statements is null, not 10. When the prerequisite is met (¿=2 comparable statements about the same fact), score 0–10 by how consistent they are. Higher score = less conflict, more consistently supporting the correct answer. (Only judge whether contradictions exist within the doc set; do not factor in relevance here.)

- 10: the two-or-more statements about the same fact are fully consistent, with no contradiction.
- 8: overall consistent; only a minor non-key wording difference that can be ignored.
- 5: slight inconsistency exists, but it does not affect reaching the correct conclusion.
- 3: clear disagreement on a key fact, potentially misleading the conclusion.
- 0: a direct factual contradiction exists (e.g., two mutually exclusive values for the same fact), undermining the set's support for the correct answer.

III. Global-Level dimensions: give ONE overall score for the WHOLE doc set

- Completeness: whether the doc set as a whole covers ALL key information points the correct answer depends on, with no obvious information gap. Coverage counts ONLY from the RELEVANT SUBSET (Relevance ¿= 1): an off-topic document (Relevance = 0) NEVER counts as covering any key information point. (This dimension measures the information GAP, so if the relevant subset covers none of the required points, Completeness = 0 — it is 0, not null.)
  - 10: fully covers all key information points the rubric lists, with no gap, sufficient to derive the correct answer.
  - 8: covers the vast majority of key points, missing only a minor piece; largely answerable.
  - 5: covers some key points but misses important information; hard to answer completely.
  - 3: only sporadically covers a few points; key information largely missing.
  - 0: covers almost no key point; impossible to answer from it.
- Density: judged in terms of LENGTH / span — whether the content that is actually useful for answering the query (the text that touches the required information points) occupies the MAJORITY of each document's length, rather than the documents being padded with large amounts of irrelevant / filler text where only a small fraction of the length is useful. Higher score = the documents are compact and on-point (high useful-content-to-length ratio); lower = the documents are long but mostly filler, with only a small useful snippet. This concerns the internal signal-to-length ratio of the documents, NOT whether a document is removable. APPLICABILITY PREREQUISITE (judge this FIRST): Density is judged ONLY over the RELEVANT SUBSET (Relevance ¿= 1), and "useful content" means ONLY the text that actually carries the query-required key-information-point content. Off-topic documents (Relevance = 0) are NOT counted at all (neither their length nor as useful content). If NO relevant document carries any key-information-point content, Density is NOT APPLICABLE: output it as null (do NOT give a score just because documents look short/clean).
  - 10: across the documents, useful content occupies the vast majority of each document's length; the documents are compact and on-point with almost no filler.
  - 8: useful content occupies most of the length in most documents; only minor filler / padding.
  - 5: useful content and irrelevant filler are roughly balanced; a fair share of the length is not useful.
  - 3: in most documents only a small fraction of the length is useful; the bulk is filler or off-topic padding.
  - 0: the documents are almost entirely filler; only a tiny sliver of the length is useful for this query.
- Reachability: whether, using ONLY this doc set and NO external knowledge, the COMPLETE reasoning chain required to reach the correct answer can be completed (every link grounded in the doc set). Only content from the RELEVANT SUBSET (Relevance ¿= 1) may ground a link; off-topic documents (Relevance = 0) contribute NOTHING to the chain. (This dimension measures derivability, so if the relevant subset grounds no link of the required key-information-point reasoning chain, Reachability = 0 — it is 0, not null.)
  - 10: every link of the reasoning chain (bridge entities, intermediate facts, final comparison, etc.) is grounded in the set; the correct answer is self-containedly derivable.
  - 8: the vast majority of links are supported by the set; only a minor link needs slight guessing.
  - 5: a key link is partially missing; some external knowledge is needed to complete the chain.
  - 3: the reasoning chain breaks in multiple places; many links lack document grounding.
  - 0: a key link is missing; it is almost impossible to derive the correct answer from the doc set alone.

Scoring example (Case)

Suppose the rubrics are:

[Relevance] (Doc-Level) "Does each document directly discuss the 1999 regular-season win-loss record of the team that won Super Bowl XXXIV (the St. Louis Rams)?"

[Authenticity] (Doc-Level) "Is each document's statement about that team's 1999 regular-season record consistent with the objective fact (13–3)?"

[Quality] (Doc-Level) "Is each document clearly structured so a reader can directly extract the 13–3 record?"

[Complementarity] (Set-Level) "Does the doc set, through complementary documents, jointly cover both key facets — 'the identity of the champion team' and 'the 1999 regular-season 13–3 record'?"

[Redundancy] (Set-Level) "Does the doc set avoid multiple documents redundantly covering the same required information point (the champion identity or the 13–3 record) without adding a new required point?"

[Conflict] (Set-Level) "Are the statements across the doc set about the Rams' 1999 regular-season record consistent, with no mutual contradiction?"

[Completeness] (Global-Level) "Does the doc set fully provide the champion team and its 1999 regular-season 13–3 record, sufficient to answer the question?"

[Density] (Global-Level) "Across the documents, does the text useful for the champion team and its 13–3 record occupy the majority of each document's length, rather than being buried in filler?"

[Reachability] (Global-Level) "Using only this doc set, can the full reasoning chain 'confirm the champion is the Rams → find their 1999 record 13–3' be completed?"

Suppose the answer requires 2 key information points: (P1) the champion is the St. Louis Rams; (P2) their 1999 regular-season record is 13–3.

Doc-Level dimensions (score each document):

- Doc A (text: "...the Rams finished 1999 at 13–3 and won Super Bowl XXXIV..."): Relevance = 10: directly and explicitly gives the champion team's 1999 record. Authenticity = 10: 13–3 fully matches the fact. Quality = 10: the record is clear and directly extractable.
- Doc B (text: about the Titans "finishing 1999 at 13–3, losing to the Rams"): Relevance = 5: discusses the same game's record, but the subject is the losing Titans rather than the champion Rams, so only indirectly relevant. Authenticity = 5: the Titans indeed went 13–3, but using it to answer "the champion's record" is misleading (both teams coincidentally went 13–3). Quality = 8: the record is clearly stated, but the reader must distinguish which team.
- Doc C (text: about a different Super Bowl XLI, some team 12–4): Relevance = 0: an entirely different edition, unrelated to this question. Authenticity = 0: makes no statement about the Rams' 1999 record. Quality = 0: unrelated to the query, contains no target information.

Set-Level dimensions (one overall score):

- Complementarity = 3: Doc A ALONE already covers both required points (P1 champion = Rams, P2 record = 13–3), so there is essentially

no division of labour; Doc B only adds non-required Titans context (not a new required point) and Doc C is unrelated — no genuine cross-document complementarity, even though coverage (Completeness) is perfect.
- Redundancy = 5: Doc A already covers both required points (P1 and P2); Doc B mainly repeats the 13–3 point (P2) from the Titans' angle without adding a NEW required point, so relative to the minimal covering set (Doc A alone) there is some redundancy; Doc C is unrelated and covers no required point.
- Conflict = null (NOT APPLICABLE): the rubric points to the RAMS' 1999 record; only Doc A actually states the Rams' record (13–3). Doc B states the TITANS' record, not the Rams', and Doc C is unrelated — so there are FEWER THAN TWO comparable statements about the same fact (the Rams' record). Since nothing could agree or contradict, Conflict is not applicable and is output as null (NOT 10).
Global-Level dimensions (one overall score):
- Completeness = 10: Doc A already gives "the Rams won + 1999 record 13–3" (both P1 and P2); the whole set has no information gap.
- Density = 8: judged ONLY over the relevant subset (Doc A and Doc B; Doc C has Relevance = 0 and is EXCLUDED — it does NOT count as filler here, since Density ignores off-topic documents). Doc A is compact and on-point; Doc B carries some extra non-essential Titans narrative, so across the RELEVANT documents the useful key-information-point text still occupies most of the length.
- Reachability = 10: Doc A alone completes the full reasoning chain "champion = Rams → record = 13–3", with no external knowledge needed.
Output format
Output ONLY a single valid JSON object (no markdown code fences, no extra commentary). The object has exactly ONE top-level field:
1. "score": the atomic scores per dimension. Doc-level dimensions use per_doc (one entry per document, with only doc_id and score); set-level / global-level dimensions use set_level (one overall score). You do NOT need to compute rubric_avg or dimension_score (those are computed by downstream code). Do NOT output any reasoning, explanation, or "reason" field — output the "score" object only.
Structure:

```
{
  "score": {
    "Relevance": {
      "level": "doc",
      "rubrics": [
        { "rubric": "<the original rubric text>",
          "per_doc": [ {"doc_id": 9, "score": 10}, {"doc_id": 18, "score": 5} ] }
      ]
    },
    "Authenticity": { "level": "doc", "rubrics": [ ... ] },
    "Quality": { "level": "doc", "rubrics": [ ... ] },
    "Complementarity": {
      "level": "set",
      "rubrics": [ { "rubric": "<the original rubric text>", "set_level": {"score": 8} } ]
    },
    "Redundancy": { "level": "set", "rubrics": [ ... ] },
    "Conflict": { "level": "set", "rubrics": [ ... ] },
    "Completeness": { "level": "set", "rubrics": [ ... ] },
    "Density": { "level": "set", "rubrics": [ ... ] },
    "Reachability": { "level": "set", "rubrics": [ ... ] }
  }
}
```

If CASE A applies (ALL documents have Relevance = 0), output the 8 non-Relevance dimensions as null, like: "Authenticity": null, "Quality": null, "Complementarity": null, "Redundancy": null, "Conflict": null, "Completeness": null, "Density": null, "Reachability": null. (Only "Relevance" is fully scored in CASE A.)
Not-applicable dimensions (independent of CASE A): even in CASE B, a relevance-grounded set/global dimension whose APPLICABILITY PREREQUISITE is not met over the RELEVANT SUBSET must be output as a bare JSON null for the WHOLE dimension (NOT a high score), while the other dimensions are still scored:
  - "Conflict": null when fewer than two RELEVANT documents make comparable statements about the same query-relevant key-information-point fact.
  - "Redundancy": null when fewer than two RELEVANT documents cover a key information point (nothing could possibly overlap).
  - "Complementarity": null when fewer than two RELEVANT documents each contribute a DIFFERENT key information point (no possible division of labour).
  - "Density": null when NO relevant document carries any key-information-point content.
(Completeness and Reachability are NEVER null in CASE B — they measure information GAP / derivability, so when the relevant subset covers nothing they are 0, not null.)
Rules:
- Output exactly ONE top-level field: "score". Do NOT output any "reason", per-item evidence, or per-item reason fields.
- Only output dimensions actually provided in the input; under each dimension, the number and order of rubrics must match the input.
- Doc-level dimensions: per_doc must cover EVERY document in the doc set. CRITICAL: doc_id MUST be exactly the id shown in square brackets before each document in the "Doc set" below (that is the document's original id, e.g. [9] → doc_id 9). Do NOT renumber the documents as 1,2,3,…; use their bracketed original ids.
- set-level / global-level dimensions: give only one overall set_level score, not per-document.
- RELEVANCE GATE: always score Relevance. If ALL documents have Relevance = 0 (CASE A), output the other 8 dimensions as null. Otherwise (CASE B) evaluate the other dimensions ONLY over the RELEVANT SUBSET (Relevance ¿= 1), and output as null any relevance-grounded dimension whose applicability prerequisite is not met.
- RELEVANCE-GROUNDED dimensions: Complementarity, Redundancy, Conflict, Density — and the coverage / derivability judged by Completeness / Reachability — are assessed ONLY over the RELEVANT SUBSET (documents with Relevance ¿= 1) and ONLY on the query-required key information points. Off-topic documents (Relevance = 0) are EXCLUDED: they never count as coverage, a distinct point, useful length, or a comparable statement, and they NEVER raise these scores. When a relevance-grounded SET-LEVEL dimension has too few relevant documents to be meaningful (see each prerequisite: Complementarity / Redundancy / Conflict / Density), output it as null — never reward it just because off-topic documents trivially "don't conflict / don't overlap / don't dilute".
- Conflict is CONDITIONAL: only score it (0–10) when at least two documents make comparable statements about the SAME query-relevant fact; if there are fewer than two such comparable statements, output "Conflict": null (do NOT output 10). Never treat "nothing to contradict" as a high Conflict score.
- Complementarity vs Completeness are DIFFERENT: Complementarity judges cross-document DIVISION OF LABOUR (if a single document covers everything by itself, Complementarity is LOW even when nothing is missing), and must NOT be lowered just because some point is missing; Completeness judges only whether any required information point is MISSING (an information gap), regardless of how many documents supply it. Do not conflate the two.

- score can only be an integer 0–10 (or null for a whole dimension under CASE A, or when a relevance-grounded dimension's applicability prerequisite is not met over the RELEVANT SUBSET — i.e. Conflict / Redundancy / Complementarity / Density); no N/A or empty per-item values allowed.

User: User query:
<QUERY>
Doc set:
<DOCS>
Rubrics (each labeled with its type and content):
<RUBRICS>
Now begin outputting the result (JSON with only the "score" object):

F.6 Rubric-based Judge Scoring Prompt: Long-form Scenario

System: You are a senior search-result quality assessor. Given a user query, a candidate document set (doc set, variable size) and a list of rubrics (each labeled with its dimension type and judgment content), you must score every rubric on a 0–10 scale, strictly grounded in the document content.
The query intent may clarify why this query was issued and what information it is trying to find, but it may also contain broader, exploratory, or irrelevant thoughts. Stay anchored on the query itself and only use the parts of the query intent that are actually relevant.
This round evaluates all 9 dimensions at once, grouped into three categories:
[A. Doc-Level dimensions] — judged for EACH individual document in the doc set: Relevance, Authenticity, Quality.
[B. Set-Level dimensions] — judged for the ENTIRE doc set as a whole; give ONE overall score, not per-document: Complementarity, Redundancy, Conflict.
[C. Global-Level dimensions] — also judged for the ENTIRE doc set as a whole; give ONE overall score, not per-document: Completeness, Density, Reachability.
Absolute prerequisite: you MUST read the documents first
- Every score must be grounded in the actual document text. Judge only based on what ACTUALLY appears in the documents; do not use your own external knowledge to fill in information the documents do not state.
- The rubric text is itself query-specific (it already contains the concrete entities, facts, and numbers of the question); the user query is also provided above for context. Judge directly against the rubric, grounded in the documents.
Evaluation procedure — RELEVANCE GATE (read carefully, follow the order)
Think step by step in this order:
1. FIRST evaluate the Relevance dimension for EVERY document in the doc set (give each document its per_doc Relevance score).
2. Define the RELEVANT SUBSET = the documents whose Relevance score ¿= 1 (i.e. documents that actually touch the query's key information point(s)). Documents with Relevance = 0 are OFF-TOPIC and are EXCLUDED from EVERY other dimension: they NEVER count as providing coverage, as a distinct information point, as useful length, or as a comparable statement, and they NEVER count toward "no redundancy" or "no conflict".
3. Then check the relevance outcome:
   - CASE A — the RELEVANT SUBSET is EMPTY (every document's Relevance score = 0): the documents this ranker selected are all useless for this question, so the other 8 dimensions (Authenticity, Quality, Complementarity, Redundancy, Conflict, Completeness, Density, Reachability) carry NO evaluation meaning. In this case DO NOT score them — output each of those 8 dimensions as null (see output format). Only the Relevance dimension is scored.
   - CASE B — the RELEVANT SUBSET is non-empty (at least ONE document has Relevance ¿= 1): evaluate the other 8 dimensions ONLY over the RELEVANT SUBSET, and apply each dimension's own APPLICABILITY PREREQUISITE below. A relevance-grounded dimension whose prerequisite is not met over the relevant subset (e.g. it needs a comparison / division of labour among relevant documents but there are not enough relevant documents, or no relevant document actually carries the key-information-point content it measures) MUST be output as null — NOT as a high score. Never reward a dimension merely because off-topic documents trivially "don't conflict / don't overlap / don't dilute".
4. Never leave Relevance itself null — Relevance is always scored.
Scoring scale (0–10)
Overall meaning (higher = better satisfies the rubric). Use the following anchors, and you MAY use intermediate integers (0–10) for in-between cases:
- 10 = Completely: fully satisfied, with sufficient detail and evidence.
- 8 = Mostly: largely satisfied, but missing some detail or support.
- 5 = Moderately: relevant content is mentioned, but key details are missing.
- 3 = Barely: not explicitly stated, only weakly inferable.
- 0 = Not at all: not addressed at all / irrelevant to the rubric / contradicts the facts.
I. Doc-Level dimensions: score EACH document individually
- Relevance: to what extent the document DIRECTLY discusses the specific information facet the rubric points to.
   - 10: the document's core content directly and fully discusses this specific question and can directly support the answer.
   - 8: directly discusses the topic, but the information is incomplete or unfocused.
   - 5: mentions the related topic only in passing, missing key information.
   - 3: barely discusses it directly, relevance is only weakly inferable.
   - 0: entirely off-topic, merely matching keywords or unrelated to the question.
- Authenticity: how well the facts stated in the document CONFORM to the reference fact given by the rubric. Only assess whether the STATEMENTS that appear are accurate; independent of relevance.
   - 10: the document explicitly states the relevant fact, fully consistent with the reference fact.
   - 8: the statement is consistent with the fact, but phrased unclearly or somewhat vaguely.
   - 5: partially correct, or the fact is right but wrapped in easily-misleading phrasing.
   - 3: only faintly touched upon; correctness is hard to judge or slightly off.
   - 0: the statement contradicts the reference fact, has an obvious factual error, or the document makes no statement related to that fact at all.
- Quality: how easy the document makes it for a reader to EXTRACT the target information, in terms of information organization and presentation. Assess presentation quality only, not factual correctness.
   - 10: clearly structured and focused; the target information is immediately apparent.
   - 8: mostly clear; the target information is extractable but mixed with some irrelevant content.
   - 5: the target information exists but is diluted by lots of irrelevant content and takes effort to find.

- 3: messy layout, high noise; the target information is very hard to extract.
- 0: empty, extremely noisy, or nearly unparseable content, OR the document is entirely off-topic and contains no target information at all.

II. Set-Level dimensions: give ONE overall score for the WHOLE doc set

- Complementarity: whether the key information points are covered through DIVISION OF LABOUR across DIFFERENT documents — i.e., different documents each contribute a DIFFERENT needed piece (one gives an overview, another the details; one supplies point P1, another supplies point P2), so that the whole is more valuable than any single document. APPLICABILITY PREREQUISITE (judge this FIRST): Complementarity is judged ONLY over the RELEVANT SUBSET (Relevance ¿= 1) and ONLY on the query-required key information points; off-topic documents (Relevance = 0) NEVER count as a contributing part. If FEWER THAN TWO relevant documents each contribute a DIFFERENT required point (there is at most one relevant document, or all relevant documents only carry the SAME single point), there is no possible cross-document division of labour on the key information points — Complementarity is NOT APPLICABLE: output it as null (NOT a high score). IMPORTANT: this dimension judges only the complementary STRUCTURE among the RELEVANT documents; it does NOT judge whether the set is complete. Do NOT lower the score merely because some required point is missing (that is Completeness's job) — an incomplete set can still be highly complementary if the points it DOES cover come from different documents. Conversely, if a SINGLE document already covers everything on its own, complementarity is LOW even though coverage may be perfect (there is no division of labour).
  - 10: strong complementarity — multiple documents clearly divide the labour, each contributing a DIFFERENT needed point, so the whole is far more valuable than any single document.
  - 8: mostly complementary — most of the covered points are supplied by different documents; only an occasional point is carried alone by one document.
  - 5: partial complementarity — there is some division of labour, but a large share of the covered content is concentrated in a single document.
  - 3: very weak complementarity — almost all useful content comes from one document, or several documents merely restate the same single point.
  - 0: no complementarity at all — only one document contributes any useful content (the others add nothing), OR a single document already covers everything by itself so no cross-document complementarity exists.
- Redundancy: judged against the MINIMAL document subset needed to cover the key information points the answer requires (the SAME key information points as Completeness). It checks whether each document brings INCREMENTAL coverage of those required information points, or merely repeats a point already covered by another document. If two or more documents cover the SAME required information point while adding no new required point, that overlap is redundancy — those documents could be dropped from the minimal covering set without losing any required point. (Example: if answering the query needs 2 information points, and both doc [1] and doc [5] only answer the SAME single point, that is redundancy.) Higher score = less redundancy (each document adds a new required point; close to a minimal covering set); lower = many documents duplicating the same point with no increment. APPLICABILITY PREREQUISITE (judge this FIRST): Redundancy is judged ONLY over the RELEVANT SUBSET (Relevance ¿= 1), against the required key information points; off-topic documents (Relevance = 0) are IGNORED and NEVER count as "a distinct point" nor improve the score. If FEWER THAN TWO relevant documents actually cover a key information point (there is nothing that could possibly overlap), Redundancy is NOT APPLICABLE: output it as null — do NOT output 10, and never reward "nothing to overlap".
  - 10: no redundancy — every document contributes a DISTINCT required information point; the set is essentially a minimal covering set with no duplicated point.
  - 8: nearly no redundancy — at most a negligible overlap; almost every document still adds a new required point.
  - 5: some redundancy — a few documents repeat an already-covered information point and add no new required point.
  - 3: heavy redundancy — many documents cover the same information point(s), giving little incremental coverage of the required points.
  - 0: severe redundancy — most documents duplicate the same one/few information points and bring no new required point (e.g., several documents all only answer the same single point).
- Conflict: whether different documents in the set are FREE OF mutual contradiction on the same factual point. APPLICABILITY PREREQUISITE (judge this FIRST): Conflict is judged ONLY over the RELEVANT SUBSET (Relevance ¿= 1); off-topic documents (Relevance = 0) never count as a comparable statement. Conflict can only be scored when the RELEVANT SUBSET contains AT LEAST TWO comparable statements about the SAME query-relevant key-information-point factual point (i.e., two or more documents each make a checkable claim about the same fact, so that agreement or contradiction is even possible). If FEWER THAN TWO comparable statements exist about the same fact (the relevant fact is asserted by at most one document, or is not asserted at all), then there is nothing that could agree or contradict — Conflict is NOT APPLICABLE. In that case DO NOT give a 0–10 score; output this dimension as null (see output format). Do NOT give a high score merely because "nothing contradicts": absence of comparable statements is null, not 10. When the prerequisite is met (¿=2 comparable statements about the same fact), score 0–10 by how consistent they are. Higher score = less conflict, more consistently supporting the correct answer. (Only judge whether contradictions exist within the doc set; do not factor in relevance here.)
  - 10: the two-or-more statements about the same fact are fully consistent, with no contradiction.
  - 8: overall consistent; only a minor non-key wording difference that can be ignored.
  - 5: slight inconsistency exists, but it does not affect reaching the correct conclusion.
  - 3: clear disagreement on a key fact, potentially misleading the conclusion.
  - 0: a direct factual contradiction exists (e.g., two mutually exclusive values for the same fact), undermining the set's support for the correct answer.

III. Global-Level dimensions: give ONE overall score for the WHOLE doc set

- Completeness: whether the doc set as a whole covers ALL key information points the correct answer depends on, with no obvious information gap. Coverage counts ONLY from the RELEVANT SUBSET (Relevance ¿= 1): an off-topic document (Relevance = 0) NEVER counts as covering any key information point. (This dimension measures the information GAP, so if the relevant subset covers none of the required points, Completeness = 0 — it is 0, not null.)
  - 10: fully covers all key information points the rubric lists, with no gap, sufficient to derive the correct answer.
  - 8: covers the vast majority of key points, missing only a minor piece; largely answerable.
  - 5: covers some key points but misses important information; hard to answer completely.
  - 3: only sporadically covers a few points; key information largely missing.
  - 0: covers almost no key point; impossible to answer from it.
- Density: judged in terms of LENGTH / span — whether the content that is actually useful for answering the query (the text that touches the required information points) occupies the MAJORITY of each document's length, rather than the documents being padded with large amounts of irrelevant / filler text where only a small fraction of the length is useful. Higher score = the documents are compact and on-point (high useful-content-to-length ratio); lower = the documents are long but mostly filler, with only a small useful snippet. This concerns the internal signal-to-length ratio of the documents, NOT whether a document is removable. APPLICABILITY PREREQUISITE (judge this FIRST): Density is judged ONLY over the RELEVANT SUBSET (Relevance ¿= 1), and "useful content" means ONLY the text that actually carries the query-required key-information-point content. Off-topic documents (Relevance = 0) are NOT counted at all (neither their length nor as useful content). If NO relevant document carries any key-information-point content, Density is NOT APPLICABLE: output it as null (do NOT give a score just because documents look short/clean).
  - 10: across the documents, useful content occupies the vast majority of each document's length; the documents are compact and on-point with almost no filler.

- 8: useful content occupies most of the length in most documents; only minor filler / padding.
- 5: useful content and irrelevant filler are roughly balanced; a fair share of the length is not useful.
- 3: in most documents only a small fraction of the length is useful; the bulk is filler or off-topic padding.
- 0: the documents are almost entirely filler; only a tiny sliver of the length is useful for this query.

- Reachability: whether, using ONLY this doc set and NO external knowledge, the COMPLETE reasoning chain required to reach the correct answer can be completed (every link grounded in the doc set). Only content from the RELEVANT SUBSET (Relevance ¿= 1) may ground a link; off-topic documents (Relevance = 0) contribute NOTHING to the chain. (This dimension measures derivability, so if the relevant subset grounds no link of the required key-information-point reasoning chain, Reachability = 0 — it is 0, not null.)
  - 10: every link of the reasoning chain (bridge entities, intermediate facts, final comparison, etc.) is grounded in the set; the correct answer is self-containedly derivable.
  - 8: the vast majority of links are supported by the set; only a minor link needs slight guessing.
  - 5: a key link is partially missing; some external knowledge is needed to complete the chain.
  - 3: the reasoning chain breaks in multiple places; many links lack document grounding.
  - 0: a key link is missing; it is almost impossible to derive the correct answer from the doc set alone.

Scoring example (Case)

Suppose the rubrics are:

[Relevance] (Doc-Level) "Does each document directly discuss the 1999 regular-season win-loss record of the team that won Super Bowl XXXIV (the St. Louis Rams)?"

[Authenticity] (Doc-Level) "Is each document's statement about that team's 1999 regular-season record consistent with the objective fact (13–3)?"

[Quality] (Doc-Level) "Is each document clearly structured so a reader can directly extract the 13–3 record?"

[Complementarity] (Set-Level) "Does the doc set, through complementary documents, jointly cover both key facets — 'the identity of the champion team' and 'the 1999 regular-season 13–3 record'?"

[Redundancy] (Set-Level) "Does the doc set avoid multiple documents redundantly covering the same required information point (the champion identity or the 13–3 record) without adding a new required point?"

[Conflict] (Set-Level) "Are the statements across the doc set about the Rams' 1999 regular-season record consistent, with no mutual contradiction?"

[Completeness] (Global-Level) "Does the doc set fully provide the champion team and its 1999 regular-season 13–3 record, sufficient to answer the question?"

[Density] (Global-Level) "Across the documents, does the text useful for the champion team and its 13–3 record occupy the majority of each document's length, rather than being buried in filler?"

[Reachability] (Global-Level) "Using only this doc set, can the full reasoning chain 'confirm the champion is the Rams → find their 1999 record 13–3' be completed?"

Suppose the answer requires 2 key information points: (P1) the champion is the St. Louis Rams; (P2) their 1999 regular-season record is 13–3.

Doc-Level dimensions (score each document):

- Doc A (text: "…the Rams finished 1999 at 13–3 and won Super Bowl XXXIV…"): Relevance = 10: directly and explicitly gives the champion team's 1999 record. Authenticity = 10: 13–3 fully matches the fact. Quality = 10: the record is clear and directly extractable.
- Doc B (text: about the Titans "finishing 1999 at 13–3, losing to the Rams"): Relevance = 5: discusses the same game's record, but the subject is the losing Titans rather than the champion Rams, so only indirectly relevant. Authenticity = 5: the Titans indeed went 13–3, but using it to answer "the champion's record" is misleading (both teams coincidentally went 13–3). Quality = 8: the record is clearly stated, but the reader must distinguish which team.
- Doc C (text: about a different Super Bowl XLI, some team 12–4): Relevance = 0: an entirely different edition, unrelated to this question. Authenticity = 0: makes no statement about the Rams' 1999 record. Quality = 0: unrelated to the query, contains no target information.

Set-Level dimensions (one overall score):

- Complementarity = 3: Doc A ALONE already covers both required points (P1 champion = Rams, P2 record = 13–3), so there is essentially no division of labour; Doc B only adds non-required Titans context (not a new required point) and Doc C is unrelated — no genuine cross-document complementarity, even though coverage (Completeness) is perfect.
- Redundancy = 5: Doc A already covers both required points (P1 and P2); Doc B mainly repeats the 13–3 point (P2) from the Titans' angle without adding a NEW required point, so relative to the minimal covering set (Doc A alone) there is some redundancy; Doc C is unrelated and covers no required point.
- Conflict = null (NOT APPLICABLE): the rubric points to the RAMS' 1999 record; only Doc A actually states the Rams' record (13–3). Doc B states the TITANS' record, not the Rams', and Doc C is unrelated — so there are FEWER THAN TWO comparable statements about the same fact (the Rams' record). Since nothing could agree or contradict, Conflict is not applicable and is output as null (NOT 10).

Global-Level dimensions (one overall score):

- Completeness = 10: Doc A already gives "the Rams won + 1999 record 13–3" (both P1 and P2); the whole set has no information gap.
- Density = 8: judged ONLY over the relevant subset (Doc A and Doc B; Doc C has Relevance = 0 and is EXCLUDED — it does NOT count as filler here, since Density ignores off-topic documents). Doc A is compact and on-point; Doc B carries some extra non-essential Titans narrative, so across the RELEVANT documents the useful key-information-point text still occupies most of the length.
- Reachability = 10: Doc A alone completes the full reasoning chain "champion = Rams → record = 13–3", with no external knowledge needed.

Output format

Output ONLY a single valid JSON object (no markdown code fences, no extra commentary). The object has exactly ONE top-level field:
1. "score": the atomic scores per dimension. Doc-level dimensions use per_doc (one entry per document, with only doc_id and score); set-level / global-level dimensions use set_level (one overall score). You do NOT need to compute rubric_avg or dimension_score (those are computed by downstream code). Do NOT output any reasoning, explanation, or "reason" field — output the "score" object only.

Structure:

```
{
  "score": {
    "Relevance": {
      "level": "doc",
      "rubrics": [
        { "rubric": "<the original rubric text>",
          "per_doc": [ {"doc_id": 9, "score": 10}, {"doc_id": 18, "score": 5} ] }
      ]
    },
    "Authenticity": { "level": "doc", "rubrics": [ ... ] },
    "Quality": { "level": "doc", "rubrics": [ ... ] },
    "Complementarity": {
      "level": "set",
```

```
      "rubrics": [ { "rubric": "<the original rubric text>", "set_level": {"score": 8} } ]
    },
    "Redundancy": { "level": "set", "rubrics": [ ... ] },
    "Conflict": { "level": "set", "rubrics": [ ... ] },
    "Completeness": { "level": "set", "rubrics": [ ... ] },
    "Density": { "level": "set", "rubrics": [ ... ] },
    "Reachability": { "level": "set", "rubrics": [ ... ] }
  }
}
```

If CASE A applies (ALL documents have Relevance = 0), output the 8 non-Relevance dimensions as null, like: "Authenticity": null, "Quality": null, "Complementarity": null, "Redundancy": null, "Conflict": null, "Completeness": null, "Density": null, "Reachability": null. (Only "Relevance" is fully scored in CASE A.)
Not-applicable dimensions (independent of CASE A): even in CASE B, a relevance-grounded set/global dimension whose APPLICABILITY PREREQUISITE is not met over the RELEVANT SUBSET must be output as a bare JSON null for the WHOLE dimension (NOT a high score), while the other dimensions are still scored:
- "Conflict": null when fewer than two RELEVANT documents make comparable statements about the same query-relevant key-information-point fact.
- "Redundancy": null when fewer than two RELEVANT documents cover a key information point (nothing could possibly overlap).
- "Complementarity": null when fewer than two RELEVANT documents each contribute a DIFFERENT key information point (no possible division of labour).
- "Density": null when NO relevant document carries any key-information-point content.

(Completeness and Reachability are NEVER null in CASE B — they measure information GAP / derivability, so when the relevant subset covers nothing they are 0, not null.)
Rules:
- Output exactly ONE top-level field: "score". Do NOT output any "reason", per-item evidence, or per-item reason fields.
- Only output dimensions actually provided in the input; under each dimension, the number and order of rubrics must match the input.
- Doc-level dimensions: per_doc must cover EVERY document in the doc set. CRITICAL: doc_id MUST be exactly the id shown in square brackets before each document in the "Doc set" below (that is the document's original id, e.g. [9] → doc_id 9). Do NOT renumber the documents as 1,2,3,...; use their bracketed original ids.
- set-level / global-level dimensions: give only one overall set_level score, not per-document.
- RELEVANCE GATE: always score Relevance. If ALL documents have Relevance = 0 (CASE A), output the other 8 dimensions as null. Otherwise (CASE B) evaluate the other dimensions ONLY over the RELEVANT SUBSET (Relevance ¿= 1), and output as null any relevance-grounded dimension whose applicability prerequisite is not met.
- RELEVANCE-GROUNDED dimensions: Complementarity, Redundancy, Conflict, Density — and the coverage / derivability judged by Completeness / Reachability — are assessed ONLY over the RELEVANT SUBSET (documents with Relevance ¿= 1) and ONLY on the query-required key information points. Off-topic documents (Relevance = 0) are EXCLUDED: they never count as coverage, a distinct point, useful length, or a comparable statement, and they NEVER raise these scores. When a relevance-grounded SET-LEVEL dimension has too few relevant documents to be meaningful (see each prerequisite: Complementarity / Redundancy / Conflict / Density), output it as null — never reward it just because off-topic documents trivially "don't conflict / don't overlap / don't dilute".
- Conflict is CONDITIONAL: only score it (0–10) when at least two documents make comparable statements about the SAME query-relevant fact; if there are fewer than two such comparable statements, output "Conflict": null (do NOT output 10). Never treat "nothing to contradict" as a high Conflict score.
- Complementarity vs Completeness are DIFFERENT: Complementarity judges cross-document DIVISION OF LABOUR (if a single document covers everything by itself, Complementarity is LOW even when nothing is missing), and must NOT be lowered just because some point is missing; Completeness judges only whether any required information point is MISSING (an information gap), regardless of how many documents supply it. Do not conflate the two.
- score can only be an integer 0–10 (or null for a whole dimension under CASE A, or when a relevance-grounded dimension's applicability prerequisite is not met over the RELEVANT SUBSET — i.e. Conflict / Redundancy / Complementarity / Density); no N/A or empty per-item values allowed.

User: User query:
<QUERY>
Query intent:
<QUERY_INTENT>
Doc set:
<DOCS>
Rubrics (each labeled with its type and content):
<RUBRICS>
Now begin outputting the result (JSON with only the "score" object):