RAG-Zeval: Enhancing RAG Responses Evaluator through End-to-End Reasoning and Ranking-Based Reinforcement Learning

Kun Li[♡*], **Yunxiang Li**[♡*], **Tianhua Zhang**[♡*], **Hongyin Luo**[♢], **Xixin Wu**[♡], **James Glass**[♢], **Helen Meng**[♡]
[♡]The Chinese University of Hong Kong, Hong Kong SAR, China
[♢]Massachusetts Institute of Technology, Cambridge MA, USA
{li.kun, yli, thzhang}@link.cuhk.edu.hk

Abstract

Robust evaluation is critical for deploying trustworthy retrieval-augmented generation (RAG) systems. However, current LLM-based evaluation frameworks predominantly rely on directly prompting resource-intensive models with complex multi-stage prompts, underutilizing models' reasoning capabilities and introducing significant computational cost. In this paper, we present RAG-Zeval (**RAG**-**Z**ero **Eval**uator), a novel end-to-end framework that formulates faithfulness and correctness evaluation of RAG systems as a rule-guided reasoning task. Our approach trains evaluators with reinforcement learning, facilitating compact models to generate comprehensive and sound assessments with detailed explanation in one-pass. We introduce a ranking-based outcome reward mechanism, using preference judgments rather than absolute scores, to address the challenge of obtaining precise pointwise reward signals. To this end, we synthesize the ranking references by generating quality-controlled responses with *zero* human annotation. Experiments demonstrate RAG-Zeval's superior performance, achieving the strongest correlation with human judgments and outperforming baselines that rely on LLMs with $10-100\times$ more parameters. Our approach also exhibits superior interpretability in response evaluation[1].

1 Introduction

Retrieval-Augmented Generation (RAG) systems (Lewis et al., 2021; Gao et al., 2024; Li et al., 2024) have become a cornerstone for building knowledge-intensive NLP applications, such as question answering and fact-checking in various domains (Zhao et al., 2025; Pipitone and Alami, 2024). By integrating external knowledge retrieval with large language models (LLMs), RAG enables more accurate and contextually relevant responses (Li et al., 2025; Asai et al., 2024), especially for queries that go beyond the static knowledge encoded in model parameters. As RAG systems are increasingly deployed in real-world scenarios, robust and comprehensive evaluation is essential to assess their performance and guide further development (Yu et al., 2025).

However, evaluating the response quality of RAG systems remains challenging due to the open-ended nature of responses, particularly when generated by LLM-based models. Traditional metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), mainly used for surface-form comparisons between phrase- or sentence-length responses and references, are often coarse-grained and fail to capture semantic fidelity or factual consistency in open-ended tasks. To overcome these limitations, recent work has explored model-based evaluation strategies, particularly leveraging LLMs as automatic judges (Gu et al., 2024). Frameworks such as RAGAS (Shahul et al., 2023) and RAG-Checker (Ru et al., 2024) have demonstrated that LLMs can provide scalable and automated assessments of metrics such as context relevance and faithfulness, thereby reducing the reliance on costly human annotation and enabling efficient large-scale evaluation.

These LLM-based approaches (Shahul et al., 2023; Ru et al., 2024) predominantly prompt LLMs to operate in pipelines with multiple isolated stages, e.g., claim decomposition and then supportiveness judgment between claims. Although showing superb performance in evaluation, they rely on large-scale LLMs with advanced capabilities (e.g., GPT-4 (OpenAI et al., 2024), Llama3-70B (Grattafiori et al., 2024)), introducing significant computational costs. On the other hand, recent studies reveal that with sufficient reasoning, remarkable problem solving competences can emerge even in compact LLMs (with <10 billion parameters) (Qi et al., 2025; DeepSeek-AI et al., 2025). Building on these

[*] Equal contribution.
[1] Code and checkpoints are available here.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 24925–24943
November 4-9, 2025

insights, we study whether compact LLMs can be transformed into robust evaluators through end-to-end reasoning chains that incorporate prerequisite analytical steps aforementioned.

In this work, we present RAG-Zeval (**RAG-Z**ero **Eval**uator), a novel framework that formulates faithfulness and correctness evaluation as a rule-guided reasoning task with zero human annotation. Our approach enables the evaluators to generate comprehensive assessments end-to-end under the instruction of predefined rules, systematically performing (1) claim decomposition, (2) evidence grounding, and (3) supportiveness judgment. Distinguished from previous multi-stage pipelines, this end-to-end evaluation ensures assessment consistency through holistic reasoning and captures the interdependence between different steps. Moreover, evaluation through generation enables us to employ Reinforcement Learning with Verifiable Rewards (RLVR) (DeepSeek-AI et al., 2025) to further enhance the evaluators, in which the complete evaluation trajectories serve as the rollouts. To overcome the challenge of acquiring precise pointwise verifiable rewards, we introduce a *ranking-based outcome reward* mechanism, which operates on more easily obtainable preference judgments instead of absolute scores (Guan et al., 2025). Recognizing that high-quality rewards in open-ended generation tasks typically require expensive human annotations (Liu et al., 2025a), we further synthesize the ranking reference using Context-Aware Decoding (Shi et al., 2023) to generate quality-controlled response candidates. Combining the above together, RAG-Zeval trains compact LLMs to achieve superior evaluation capabilities with *zero* human annotation.

We assess RAG-Zeval on both faithfulness and correctness benchmarks to analyze its performance in deriving interpretable and reliable evaluations. Experimental results demonstrate that RAG-Zeval achieves strong alignment with human judgments, maintaining transparent and interpretable decision-making through its rule-guided reasoning process.

2 Related Work

With the rapid advancement of Retrieval-Augmented Generation (RAG) systems (Fan et al., 2024; Li et al., 2025), effective and robust evaluation methods beyond traditional metrics have become increasingly important.

A significant line of work evaluates the retrieval and generation components separately. For retrieval, traditional information retrieval metrics such as precision, recall, MRR, and MAP are widely used (Yu et al., 2024; Tang and Yang, 2024). For the generation component, metrics like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2020) are commonly used, alongside human evaluation.

Recent research on the evaluation of RAG systems has moved beyond traditional component-wise metrics, proposing a variety of frameworks that leverage large language models (LLMs) as evaluators. LLM-based evaluation frameworks such as TruLens (Ferrara et al., 2024) and ARES (Saad-Falcon et al., 2024) adopt direct prompting to score responses without decomposing them into individual claims. Other approaches, including RAGAS (Shahul et al., 2023), RAG-Checker (Ru et al., 2024), and OpenEval (Ispas et al., 2025), introduce claim-level decomposition, enabling LLMs to assess the faithfulness and correctness of each factual statement for finer-grained and more interpretable evaluation.

Despite these advances, most current LLM-based evaluation frameworks rely on direct prompting of large, resource-intensive models, often involve complex multi-stage prompting, and treat LLMs as black-box scorers without fully leveraging their reasoning abilities. Recent progress in reinforcement learning and reward modeling, such as Deepseek-R1 (DeepSeek-AI et al., 2025) and generative reward modeling (GRM) (Liu et al., 2025a), demonstrates that rule-driven, interpretable evaluators trained via rule-based RL can provide more transparent and scalable assessments with stronger reasoning ability. These developments motivate our approach to construct RAG evaluators using similar RL-based, rule-guided techniques.

3 Methodology

3.1 Problem Formulation

The majority of evaluation tools for assessing the response quality of RAG systems adopt a claim-based paradigm (Ru et al., 2024; Shahul et al., 2023; Ispas et al., 2025; Manakul et al., 2023). In this paradigm, the responses are decomposed into individual claims, each declarative sentence that conveys an atomic piece of information. The claims are then evaluated for supportiveness—the degree to which they are grounded on the provided reference context (e.g., ground-truth answer for

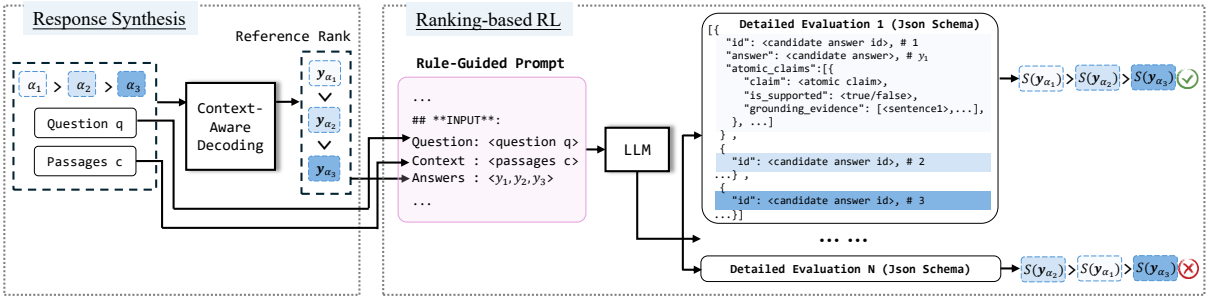

Figure 1: An overview of RAG-Zeval. We synthesize training data using Context-Aware Decoding, generating $\boldsymbol{y}_{\boldsymbol{\alpha}_i}$ with α_i. The complete prompt is presented in Fig.5. The ground-truth ranking of $\boldsymbol{y}$'s depends on the value of α.

correctness or retrieved passages for faithfulness).

In a RAG setting, a response is considered 1) correct if the ground-truth answer supports its claims, or 2) faithful if the retrieved passage supports its claims. Based on this, following prior work (Ru et al., 2024; Ispas et al., 2025; Shahul et al., 2023), given a response, we define its ***correctness*** as

$$\frac{\text{\# Claims supported by ground-truth answer}}{\text{\# Claims}}, \quad (1)$$

and ***faithfulness*** as

$$\frac{\text{\# Claims supported by retrieved passage}}{\text{\# Claims}}, \quad (2)$$

indicating the precision rate of claims in the response that are supported by the ground-truth answer and the retrieved passage respectively. **These formulations consolidate the evaluation paradigms for both correctness and faithfulness, allowing for the development of a unified evaluator that assesses the two quality dimensions**. Therefore, our approach is able to assess responses in correctness and faithfulness with different reference (i.e., ground-truth answers for correctness and retrieved passages for faithfulness evaluation).

3.2 Prompting for Rule-Guided Reasoning

Different from previous claimed-based work, which runs in a multi-stage pipeline, we develop a novel approach for end-to-end claim-based evaluation, through *generation* of complete evaluation trajectories, guided by our defined rules.

To this end, we adopt the prompt shown in Fig. 5, which elaborates the rules and format that the generation should follow. In detail, given a question $\boldsymbol{q}$ and the reference $\boldsymbol{c}$, and the set of responses to evaluate $\{\boldsymbol{y}\}$, LLM is prompted to give a comprehensive evaluation process—decomposing a response into claims, and then determining those claims' supportiveness as well as finding the grounding evidence in the reference. In addition, the generation is required to represent the evaluation process in a JSON format. After parsing the generated JSON-formatted string into a Python list object using `json.loads`, we can readily extract the intermediate results (e.g., # claims (un)supported by the reference) by accessing the resulting list object.

Casting the evaluation process into generation of evaluation trajectory not only streamlines the pipeline, but also facilitates further finetuning of the model.

3.3 Reinforcement Learning with Ranking Objective

Finetuning models with valid evaluation trajectories as outlined in §3.2 presents a non-trivial challenge due to the prohibitive cost of manual annotation—particularly for claim decomposition and supportiveness judgment. To address this, we use Reinforcement Learning with Verifiable Rewards (RLVR) (DeepSeek-AI et al., 2025) to finetune our model, bypassing the need for annotation of the whole trajectories.

Nonetheless, it remains necessary to curate the data labeled with the final evaluation result for the outcome rewards. A naive way would be to annotate the score according to Eq.1 or 2. However, this way still relies on the claim decomposition and supportiveness judgment as the intermediate results. To circumvent this, our reinforcement learning method introduces a novel optimization paradigm that trains the model to perform relative ranking of responses based on their degree of supportiveness w.r.t. the reference, rather than predicting absolute scores. Specifically, first, given question $\boldsymbol{q}$ and reference $\boldsymbol{c}$, we synthesize a set of responses $\{\boldsymbol{y}\}$ with controlled groundness degree w.r.t. $\boldsymbol{c}$ (§3.3.1), which varies across $\{\boldsymbol{y}\}$. During this process, the ground-truth rank of $\{\boldsymbol{y}\}$ can be

obtained naturally. Subsequently, based on $\boldsymbol{q}$ and $\boldsymbol{c}$, we adopt $\{\boldsymbol{y}\}$ as the responses to rank and apply RL to reinforce the model's ranking ability by advancing the generated evaluation trajectories (§3.3.2). Note that the RL objective is to rank the responses instead of predicting their exact scores. For the training, this can mitigate the adverse effects of bias introduced during data synthes.

3.3.1 Responses Synthesis with Ranking Relation

With Context-Aware Decoding (Li et al., 2022; Shi et al., 2023) , the i-th token of a response y is sampled as

$$y_i \sim softmax[(1+\alpha)Logit_{LLM}(* \mid \boldsymbol{q}, \boldsymbol{c}, \boldsymbol{y_{<i}}) - \alpha Logit_{LLM}(* \mid \boldsymbol{q}, \boldsymbol{y_{<i}})]. \tag{3}$$

$Logit_{LLM}(* \mid \boldsymbol{q}, \boldsymbol{y_{<i}})$ denotes the logits of $\boldsymbol{y_i}$ predicted by an LLM with the input of $\boldsymbol{q}$ and $\boldsymbol{y_{<i}}$. The weight α controls the extent to which the generation of y_i is conditioned on $\boldsymbol{c}$ (which is the passage in this case), and a larger one translates into $\boldsymbol{y}$ that is more reference-conditioned. Note that α can be negative and $\alpha < -1$ leads to reference-resistant generation of $\boldsymbol{y}$ (App. A.1).

For each question, we synthesize a set of responses $\{\boldsymbol{y_{a_i}}\}$ with different degrees of groundness by varying α. The ground-truth ranking of these responses can be obtained naturally as

$$\forall \alpha_i, \alpha_j \in \mathbb{R}, \quad \alpha_i > \alpha_j \implies \boldsymbol{y}_{\alpha_i} \succ \boldsymbol{y}_{\alpha_j}. \tag{4}$$

For implementation, $Logit_{LLM}(* \mid \boldsymbol{q}, \boldsymbol{c}, \boldsymbol{y_{<i}})$ and $Logit_{LLM}(* \mid \boldsymbol{q}, \boldsymbol{y_{<i}})$ are modeled using in-context learning. We use a third-party LLM for sampling responses prior to the RL stage.

3.3.2 Reinforcement Learning with Verifiable Rewards

We fine-tune the model using reinforcement learning with verifiable rewards. Particularly, we adopt Group Relative Policy Optimization (GRPO, Shao et al., 2024) with rule-based outcome rewards. During rolling out, with the question $\boldsymbol{q}$, the reference $\boldsymbol{c}$, and the set of synthesized responses $\{\boldsymbol{y_{a_i}}\}$ as input, the model generates complete evaluation trajectories according to the rules specified in the prompt.

Reward Design We define three types of rewards, including format reward, evidence reward, and accuracy reward. The rewards for a rollout of evaluation trajectory are defined as follows.

- **Format reward** assesses the completeness of the evaluation trajectory. r_{f} is **0** if the string of evaluation trajectory satisfies all the following requirements: 1) it can be parsed into a Python `List` object using `json.loads`; 2) the items in the list correspond exactly to the set of responses to evaluate; 3) each item within the `List` is a `Dict` object containing all required fields as specified in the prompt (the circled region in Fig.5); 4) each supported claim has at least one evidence. Otherwise, a penalty of $-\mathbf{0.5}$ is applied.

- **Evidence reward** measures how verbatim extracted evidence spans are cited from the reference. The reward for each span is defined as the length of its longest common substring with the reference text, normalized by the span's length[2]. An evidence span of length less than 10 receives reward **0**. The evidence reward of an evaluation trajectory r_{e} is the average over all evidence spans in the trajectory.

- **Accuracy reward** evaluates whether the ranking based on the evaluation scores inferred by the model is correct. The evaluation score is derived as $S(\boldsymbol{y}) = \frac{\text{\# Claims of } \boldsymbol{y} \text{ supported by } \boldsymbol{c}}{\text{\# Claims of } \boldsymbol{y}}$.The accuracy reward r_{a} is **1** if the ranking aligns with the ground-truth ranking, or **0** otherwise. Formally,

$$r_{\text{a}} = \begin{cases} 1, \text{ if } S(\boldsymbol{y}_{\alpha_i}) > S(\boldsymbol{y}_{\alpha_j}), \forall\ \boldsymbol{y}_{\alpha_i}, \boldsymbol{y}_{\alpha_j} \in \{\boldsymbol{y}\} \text{ and } \boldsymbol{y}_{\alpha_i} \succ \boldsymbol{y}_{\alpha_j} \\ 0, \text{ otherwise} \end{cases} \tag{5}$$

The intermediate results required for obtaining the reward a rollout can be accessed by visiting the object obtained after parsing. In particular, for each response, we apply Python operations to enumerate all entries in its `atomic_claims` list and verify their `is_supported` values. The JSON-formatted output demonstrates superior precision in results extraction, compared to traditional regular expression-based approaches.

Taking together above three rewards, the combined reward r for a rollout is

$$r = \begin{cases} 1 + 0.5 * r_{\text{e}}, & \text{if } r_{\text{f}} = 0 \text{ and } r_{\text{a}} = 1, \\ 0, & \text{if } r_{\text{f}} = 0 \text{ and } r_{\text{a}} = 0, \\ -0.5, & \text{otherwise.} \end{cases} \tag{6}$$

[2]The length of a sequence is computed as its total token count

The reward function encourages the model to rank the responses more accurately through optimizing the evaluation trajectories.

Curriculum Learning Intuitively, it is more challenging to rank a larger set of responses. In the spirit of curriculum learning (Bengio et al., 2009; Narvekar et al., 2020), to facilitate smooth and incremental learning, we gradually escalate the complexity of the ranking task by increasing the number of responses to evaluate, as the RL training process advances.

4 Experiment Settings

4.1 Benchmarks and Metrics

Faithfulness We assess the faithfulness judgment performance of different evaluation approaches on WikiEval dataset (Shahul et al., 2023), which contains question-context-answer triples with human-annotated judgments. The questions are formulated from 50 Wikipedia pages, and for each question, ChatGPT generates two answers: one with Wikipedia context and one without. Two human annotators then judge which answer is more faithful to the source, reaching 95% agreement.

For each WikiEval instance, the evaluators are required to identify the more faithful answer between two candidates. Evaluator performance is then measured as the percentage of cases where the evaluators' preference aligns with the human annotators' judgment (Shahul et al., 2023; Ispas et al., 2025). We follow Shahul et al. (2023) to handle possible *ties* with three scenarios (see App. A.3 for more details):

- **Best-Case**: Measures the frequency of evaluators assigning greater or equal faithfulness scores to good answers over poor ones.
- **Worst-Case**: Computes the frequency of strictly greater faithfulness scores assigned to good answers.
- **Middle-Case**: Adopts ternary scoring with a partial point of 0.5 for ties.

Correctness To assess different correctness evaluation approaches, we use the Meta Evaluation Dataset constructed by Ru et al. (2024). The dataset contains 280 instances from 10 domains. Each instance includes a question, the ground-truth answer, and a pair of responses generated by two RAG systems[3]. Two human annotators assess the responses, assigning preference labels from five relative choices: significantly better, slightly better, tie, slightly worse and significantly worse. We adopt human-annotated correctness preferences as the references to benchmark evaluation methods. Following Ru et al. (2024), we convert the human-annotated correctness preference labels (five relative choices) into a numerical score difference for each response pair, i.e., $h_i = H(r_i^2) - H(r_i^1) \in \{-2, -1, 0, 1, 2\}$. A normalized score difference is computed as $e_i = f(E(r_i^2) - E(r_i^1))$ for each evaluation approach, where $E(\cdot)$ is the correctness score measured by the evaluator and $f(\cdot)$ is a linear normalization function. To assess the performance of different evaluation methods, we compute three correlation coefficients between human judgments h_i and system scores e_i: Pearson's r, Spearman's ρ, and Kendall's τ.

4.2 Implementation Details

Responses Synthesis We use the corpus of Natural Question (Kwiatkowski et al., 2019) to synthesize the responses, where each question is accompanied by a grounding passage. 5,500 instances are selected randomly for response synthesis. For each $\alpha \in \{0, -0.5, -1, -1.4\}$, we synthesize a response according to Eq.3, using `Qwen2.5-7B-Instruct` (Qwen, 2024). See App. A.1 for more details.

Training We fine-tune our model based on `Qwen2.5-7B-Instruct` and `Llama-3.1-8B-Instruct` respecetively. For RL training, the sample number is 8 and temperature is 1 during rollout. The KL coefficient in the learning objective is 0.015. We train the model for a total of 2 epochs. To achieve curriculum learning for RL, we use 3 candidate responses for ranking in the first epoch and increase this to 4 in the second epoch. More details can be found in App. A.2.

Inference For test instances on both datasets described in §4.1, similar to rolling out at training stage, the evaluator model takes as input the question, the reference text (ground-truth answer for correctness and retrieved passage for faithfulness), and two candidate responses; During the generation of evaluation trajectory, we use nucleus sampling (Holtzman et al., 2019) with $p = 0.9$ and temperature $= 0.1$. For those generated sequence that fails

[3] https://github.com/amazon-science/RAGChecker/blob/main/data/meta_evaluation/

Method	*Model (-Instruct)*	*Best*	*Middle*	*Worst*
BLEU	–	0.860	0.860	0.860
RougL	–	0.900	0.900	0.900
BERTScore	–	0.900	0.900	0.900
ARES	llama-70b	1.000	0.920	0.840
	qwen-72b	1.000	0.928	0.856
	gpt-4o	1.000	0.956	0.912
TruLens	llama-70b	1.000	0.860	0.720
	qwen-72b	0.984	0.830	0.676
	gpt-4o	1.000	0.940	0.900
RAGAS	llama-70b	0.960	0.910	0.860
	qwen-72b	0.960	0.922	0.884
	gpt-4o	0.980	0.940	0.900
RAG-Checker	llama-70b	1.000	0.962	0.924
	qwen-72b	0.976	0.936	0.896
	gpt-4o	0.973	0.933	0.893
OpenEval*	llama-70b	0.960	0.950	0.940
SFT	qwen-72b	0.828	0.828	0.828
RAG-Zeval w/o RL	llama-70b	0.980	0.960	0.927
	qwen-72b	0.993	0.957	0.883
	gpt-4o	0.987	0.970	0.953
	qwen-7b	0.932	0.930	0.858
	llama-8b	0.620	0.590	0.560
RAG-Zeval w/ RL	qwen-7b	**1.000**†	**0.992**†	**0.984**†
	llama-8b	1.000	0.987	0.973

Table 1: Performance on **faithfulness** evaluation. We assess different methods using `Llama3.1-70B-Instruct`, `Qwen2.5-70B-Instruct`, `GPT-4o` and/or `Qwen2.5-7B-Instruct`. Non-GPT results are averaged over five trials to mitigate randomness. Due to API cost, we ran `GPT-4o` three times for each method. We cite results of OpenEval from the original paper (Ispas et al., 2025). † indicates the result is statistically significant at the level of 0.01.

to parse, we utilize regular expressions to extract the required results. The correctness/faithfulness score for a response $\boldsymbol{y}$ is computed as $S(\boldsymbol{y})$ (Eq.5, see Fig.7 for an example).

4.3 Baselines

We compare our approach with a comprehensive set of baseline evaluation methods, including non-LLM based and LLM-based paradigms. For non-LLM based methods, we report BLEU (Papineni et al., 2002) and ROUGE-L (Lin, 2004) as representative n-gram based metrics, as well as BERTScore (Zhang et al., 2020) for embedding-based metric. For LLM-based evaluation, we include recent frameworks that all use iterative prompting with large language models as evaluators. ARES (Saad-Falcon et al., 2024) and TruLens (Ferrara et al., 2024) are non-claim-based, directly prompting the LLM for overall or aspect-based scores. RAGAS (Shahul et al., 2023), RAG-Checker (Ru et al., 2024), and OpenEval (Ispas et al., 2025) are claim-based, decomposing responses into factual claims for individual assessment. All LLM-based baselines are re-implemented with `Llama3.1-70B-Instruct`, `Qwen2.5-70B-Instruct`, `GPT-4o` as the evaluator backbone. In addition, we consider standard SFT, which directly fine-tunes `Qwen2.5-7B-Instruct` to replicate the relative ranking of responses, using the same synthetic data described in Section 3.3.2 (see App. A.4 for more details).

Method	*Model (-Instruct)*	*Pearson*	*Spearman*	*Kendall*
BLEU	–	0.302	0.305	0.236
RougL	–	0.395	0.428	0.335
BERTScore	–	0.350	0.437	0.341
ARES	llama-70b	0.350	0.328	0.296
	qwen-72b	0.423	0.396	0.360
	gpt-4o	0.382	0.370	0.333
TruLens	llama-70b	0.428	0.453	0.366
	qwen-72b	0.428	0.446	0.360
	gpt-4o	0.396	0.390	0.312
RAGAS	embedding	0.411	0.432	0.283
RAG-Checker	llama-70b	0.463	0.425	0.337
	qwen-72b	0.495	0.465	0.375
	gpt-4o	0.499	0.459	0.369
SFT	qwen-72b	0.359	0.350	0.320
RAG-Zeval w/o RL	llama-70b	0.492	0.443	0.351
	qwen-72b	0.521†	0.482†	0.388†
	gpt-4o	**0.585**†	**0.554**†	**0.452**†
	qwen-7b	0.427	0.367	0.312
	llama-8b	0.370	0.342	0.280
RAG-Zeval w/ RL	qwen-7b	0.501	0.452	0.354
	llama-8b	0.498	0.435	0.342

Table 2: Performance on **correctness** evaluation. Correlation between different methods and human judgments are reported. We assess RAGAS (Shahul et al., 2023) with `Text-Embedding-Ada-002` model (Neelakantan et al., 2022) following the original setting. Other settings are the same as Tab.1. Following Ru et al. (2024), we only show the metric with the best correlation for each baseline framework. See more details in App. A.4. † indicates the result is statistically significant at the level of 0.01.

5 Main Experiments

Comparison with Baselines Table 1 and 2 present the performances of `RAG-Zeval` and baseline evaluators. Generally, the claim-based methods outperform non-claim-based ones. For both benchmarks, `RAG-Zeval` has the strongest correlation with human preference in terms of almost all metrics. `RAG-Zeval` shows elevating performance with larger backbone models. On the other hand, even with compact architectures (7 or 8 billion param-

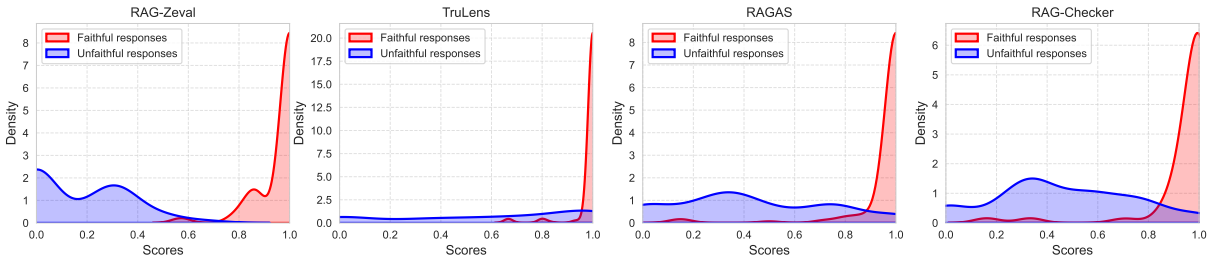

Figure 2: The density distribution of the scores assigned by the faithfulness evaluators.The distribution of the faithful and unfaithful responses are marked with red and blue, respectively. TruLens, RAGAS and RAG-Checker are all implemented with `Qwen2.5-72B-Instruct` as the backbone LLM. RAG-Zeval is obtained by finetuning `Qwen2.5-7B-Instruct` with RL.

eters), `RAG-Zeval` w/ RL demonstrates superior performance over most baselines built on large-scale LLMs with 10-100× more parameters. This result validates the effectiveness of our approach for enhancing evaluation capabilities in compact LLMs.

For in-depth comparison, Fig. 2 visualizes the distribution of scores assigned by `RAG-Zeval` w/ RL (qwen-7b) and some baselines that give numerical (instead of categorical) predictions for faithfulness evaluation, where we can see the distribution of faithful and unfaithful responses[4]. While TruLens has the most concentrated distribution near 1 for faithful responses, its distribution for unfaithful responses disperses evenly across the X-axis, indicating its inability to distinguish unfaithful responses. For faithful responses, `RAG-Zeval`, RAG-Checker and RAGAS demonstrate similar distributional shapes, particularly showing comparable peakedness near 1. However, `RAG-Zeval` shows superior discriminative capacity, maintaining clear separation between faithful and unfaithful response distributions.

Comparison with Ablated Variants As shown in Tab. 1 and 2, SFT solely on the ground-truth ranking exhibits the worst performances among the LLM-based methods. In contrast, even without further training, the non-RL version of our approach maintains robust evaluation performance. This suggests the significance of intermediate reasoning. `RAG-Zeval`, which generates the complete reasoning trajectories for evaluation, can effectively harness the LLMs' reasoning capabilities to achieve superior performance. With the identical backbone models, `RAG-Zeval` w/ RL significantly outperforms the counterpart w/o RL. It can be inferred that the reasoning ability for evaluation has been enhanced through RL. More discussion on this is in § 6.1 and § 6.3.

More experiment results on different backbone models and data splits can be found in App. B.

6 Analysis

In this section, the following problems are discussed: 1) How reinforcement learning stimulates the model's evaluation ability. 2) What is the effect of the task complexity represented by training objective and data. 3) What is the effectiveness of the rule-guided reasoning. Throughout this section, we utilize the correctness benchmark for our evaluation due to its greater challenge, and use `Qwen2.5-7B-Instruct` as the backbone model.

6.1 Self-Evolution of Evaluation Abilities

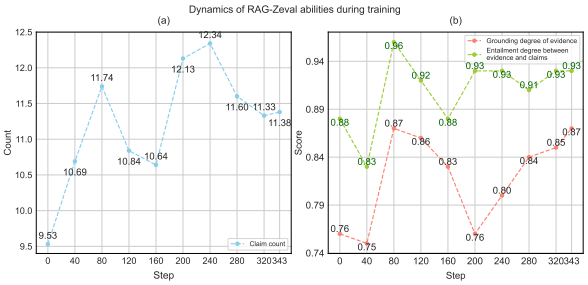

Figure 3: (a) shows the changes of decomposed claim count, while (b) presents the evolution of abilities of evidence extraction and supportiveness judgment throughout the RL training process. The statistics are based on the rollout samples during training.

To investigate how the model abilities evolve over the RL training process, we continuously monitor the model's behaviors in claim decomposition, supportiveness judgment, and evidence extraction. Their dynamics is plotted in Fig. 3.

[4]We do not use the correctness evaluation benchmark here, as the human annotators only provide relative assessment (e.g., preference ranking) rather than absolute categorical judgments (correct/incorrect labels).

Objective	*Pearson*	*Spearman*	*Kendall*
Ranking (Ours)	0.501†	0.452†	0.354†
Predicting the best	0.406	0.393	0.311

Table 3: Comparison of two training objectives. The results are obtained by averaging across 5 runs.

Data configuration	*Pearson*	*Spearman*	*Kendall*
Curriculum learning (first 3 and then 4 responses)	0.501†	0.452†	0.354†
Static (3 responses)	0.457	0.433	0.339
Static (4 responses)	0.450	0.402	0.314

Table 4: Comparison of different data configuration. The results are obtained by averaging across 5 runs.

The blue line depicts the average number of claims decomposed from a response, which initially exhibits a sharp increase and ends at a stable level. As finer-grained claim decomposition enables more discriminative comparisons among candidate responses, the model learns increasingly comprehensive claim decomposition for enhanced ranking performance. This is evidenced by the case study in Fig. 6 and 7, in which the trained checkpoint (at step 343) provides a more comprehensive decomposition, whereas some claims by the untrained checkpoint (at step 0) amalgamate atomic claims that should have been addressed separately.

For each supported claim generated by the model, we quantify the degree of textual entailment between its corresponding evidence and the claim itself, using AlignScore (Zha et al., 2023)[5]. For each extracted evidence, we measure its grounding degree in terms of the normalized length of its longest common substring (similar to evidence reward definition in §3.3.2). The entailment (green line) degrees and grounding (red line) both experience a notable growth, implying that the model's capabilities of supportiveness judgment and evidence extraction get improved through the RL training.

Overall, *reinforcement learning effectively incentivizes the development of reasoning capabilities essential for responses evaluation*, consequently improving the final evaluation performance.

[5]The evidence is input as `context`, and the claims is as `claim`. `context` and `claim` are two arguments for AlignScore which measures how likely `context` would entails `claim`.

6.2 Effect of Task Complexity

Ranking-based Objective We simplify the ranking-based accuracy reward (Eq. 5) as

$$r_a = \begin{cases} 1, \text{ if } & S(\boldsymbol{y}_{\alpha_i}) > S(\boldsymbol{y}_{\alpha_j}), \alpha_i = max\{\alpha\}, \\ & \forall \boldsymbol{y}_{\alpha_j} \in \{\boldsymbol{y}\} \text{ and } \boldsymbol{y}_{\alpha_i} \neq \boldsymbol{y}_{\alpha_j} \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

indicating that an accuracy reward of $\mathbf{1}$ is earned if the model assigns the highest evaluation score to $\boldsymbol{y}$ with the largest α value. This simplified formulation, similar to the one used by Liu et al. (2025b), does not require an correct ranking over the entire set of responses to evaluate. The comparative results between the two formulation are shown in Tab. 3. The model trained with the simplified accuracy reward suffers a notable performance drop. This implies that *reducing task complexity may diminish the incentive for the model to develop enhanced evaluation capabilities*. Because a complete ranking of all responses requires a more granular and discriminative assessment than merely predicting the top-ranked one.

Curriculum Learning During the RL training, we organize the training data in a way that the complexity of the ranking task escalates as the training advances. To study the effect of this practice, we also train models with the following two static data organization—the training instance across all epochs consistently contains 3 or 4 responses for ranking. As illustrated in Tab. 4, the curriculum learning-based configuration has the best performance. Its improvement over the static one with 3 responses further corroborates above finding that increased task complexity helps ability acquisition. On the other hand, the static configurations with 4 response performs worst. We found it earns a much lower average combined reward than the curriculum learning-based configuration in the first epoch (seen in App.C). Employing overly challenging task objective in the initial training stage may suppress model learning, as it is less possible to find a valid rollout and the model then hardly receives positive feedback during the training.

6.3 Effectiveness of Rule-Guided Reasoning

Results in § 5 demonstrates that `RAG-Zeval` outperforms direct SFT on ground-truth ranking of responses. To better illustrate the significance of our rule-guided reasoning, we further introduce an intermediate variant between the above two methods– remove the requirement to provide supporting evi-

Method	Generation	*Pearson*	*Spearman*	*Kendall*
Ours	Complete reasoning trajectories specified in Fig.5	$0.501^{\dagger}$	$0.452^{\dagger}$	$0.354^{\dagger}$
w/o evidence	Reasoning trajectories without evidence and analysis	0.489	0.452	0.353
SFT	Only response ranking result	0.359	0.350	0.320

Table 5: Results of methods with different evaluation pattern, obtained by averaging across 5 runs.

dence and corresponding analysis, while maintaining all other settings consistent with `RAG-Zeval`.

As shown in Tab. 5, their performances are positively correlated with the level of detail of their generation, which substantiates the advantage of the rule-guided reasoning. Analogues to Chain-of-Thought (Wei et al., 2022), `RAG-Zeval` benefits from the stepwise reasoning in its evaluation trajectories. Also, detailed evaluation processes offer better interpretability behind the model predictions.

6.4 Case Study

In Fig. 6 and 7, `RAG-Zeval` w/ RL and `RAG-Zeval` w/o RL generate evaluations for the same input. For this question "*price of PS3 when it first came out*", human annotators judge response B as significantly better than response A. `RAG-Zeval` w/o RL assigns response A 1 point, while response B $1/3$ point, resulting in an incorrect ranking $A \succ B$, which contradicts human preference. Additionally, its final claim is a verbatim copy of the original answer sentences, failing to perform atomic claim decomposition. In contrast, `RAG-Zeval` assigns response A 0 point and response B $3/4$ point, owing to its finer-grained claim decomposition and more accurate supportiveness judgments, ultimately yielding a correct ranking that aligns with human evaluation.

7 Conclusion

In this work, we introduce `RAG-Zeval`, a novel evaluation framework that performs end-to-end, interpretable assessment of RAG system responses. Our approach significantly improves compact LLM-based evaluators via reinforcement learning with a novel ranking-based objective, bypassing the requirement for human-annotated data. Through comprehensive experiments on benchmarks of faithfulness and correctness evaluation, we demonstrate that our approach achieves strong alignment with human judgments, outperforming current large-scale LLM-based baselines while maintaining a much smaller model scale. The result highlights the potential of compact, reasoning-driven evaluators for scalable and transparent RAG evaluation.

Acknowledgement

This study was supported by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd., a CUHK-led InnoCentre under the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government.

Limitations

This work has several limitations that point to avenues for future improvement. Although our approach employs smaller models compared to direct use of large LLMs, the RL training process still requires considerable computational resources and access to high-performance hardware, which may not be available to all researchers. In addition, our current experiments are primarily conducted in English and on general-domain datasets; the generalizability of the evaluator to other languages or specific domains remains to be explored. Further validation on multilingual and domain-specific benchmarks would strengthen the robustness and applicability of our method.

Additionally, our experiments run on static datasets, which may not capture real-world dynamic interactions well (e.g., adversarial inputs, evolving user preferences). Further investigation of its performance in real-world environments is essential prior to deployment, to ensure unbiased and accurate judgments.

Ethical Considerations

Our RL-based RAG evaluation framework also raises several ethical considerations. The computational requirements, though reduced compared to large LLMs, may still create barriers for less-resourced groups, potentially exacerbating inequities in access to advanced evaluation tools. Moreover, automated evaluation should not be viewed as a substitute for human oversight, especially in high-stakes or sensitive applications, as it

may overlook nuanced ethical or contextual factors. Besides, if the synthetic or training data used for evaluator construction contains biases or unrepresentative patterns, these biases may be propagated in the evaluation results. Responsible deployment requires ongoing attention to these issues and a commitment to transparency and fairness

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA. Association for Computing Machinery.

Baolong Bi, Shenghua Liu, Yiwei Wang, Yilong Xu, Junfeng Fang, Lingrui Mei, and Xueqi Cheng. 2025. Parameters vs. context: Fine-grained control of knowledge reliance in language models. *CoRR*, abs/2503.15888.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. *Preprint*, arXiv:2405.06211.

J. Ferrara, Ethan-Tonic, and O. M. Ozturk. 2024. The rag triad. `https://www.trulens.org/trulens_eval/core_concepts_rag_triad/`.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on llm-as-a-judge. *ArXiv*, abs/2411.15594.

Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. 2025. rstar-math: Small llms can master math reasoning with self-evolved deep thinking. *Preprint*, arXiv:2501.04519.

Rujun Han, Peng Qi, Yuhao Zhang, Lan Liu, Juliette Burger, William Yang Wang, Zhiheng Huang, Bing Xiang, and Dan Roth. 2023. RobustQA: Benchmarking the robustness of domain adaptation for open-domain question answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4294–4311, Toronto, Canada. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *ArXiv*, abs/1904.09751.

Alex-Răzvan Ispas, Charles-Elie Simon, Fabien Caspani, and Vincent Guigue. 2025. Towards lighter and robust evaluation for retrieval augmented generation. *The Next Frontier in Reliable AI": Workshop on ICLR 2025*, abs/2503.16161.

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. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Kun Li, Tianhua Zhang, Yunxiang Li, Hongyin Luo, Abdalla Moustafa, Xixin Wu, James Glass, and Helen Meng. 2025. Generate, discriminate, evolve: Enhancing context faithfulness via fine-grained sentence-level self-evolution. *Preprint*, arXiv:2503.01695.

Kun Li, Tianhua Zhang, Xixin Wu, Hongyin Luo, James Glass, and Helen Meng. 2024. Decoding on graphs: Faithful and sound reasoning on knowledge graphs through generation of well-formed chains. *Preprint*, arXiv:2410.18415.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2022. Contrastive decoding: Open-ended text generation as optimization. In *Annual Meeting of the Association for Computational Linguistics*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. 2025a. Inference-time scaling for generalist reward modeling. *Preprint*, arXiv:2504.02495.

Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. 2025b. Inference-time scaling for generalist reward modeling.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E. Taylor, and Peter Stone. 2020. Curriculum learning for reinforcement learning domains: a framework and survey. 21(1).

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, and 6 others. 2022. Text and code embeddings by contrastive pre-training. *Preprint*, arXiv:2201.10005.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Nicholas Pipitone and Ghita Houir Alami. 2024. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain. *ArXiv*, abs/2408.10343.

Zhenting Qi, Mingyuan MA, Jiahang Xu, Li Lyna Zhang, Fan Yang, and Mao Yang. 2025. Mutual reasoning makes smaller LLMs stronger problem-solver. In *The Thirteenth International Conference on Learning Representations*.

Qwen. 2024. Qwen2.5: A party of foundation models.

Sara Rosenthal, Avirup Sil, Radu Florian, and Salim Roukos. 2025. CLAPnq: Cohesive long-form answers from passages in natural questions for RAG systems. *Transactions of the Association for Computational Linguistics*, 13:53–72.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. *NeurIPS*.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354, Mexico City, Mexico. Association for Computational Linguistics.

ES Shahul, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2023. Ragas: Automated evaluation of retrieval augmented generation. In *Conference of the European Chapter of the Association for Computational Linguistics*.

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

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Yih. 2023. Trusting your evidence: Hallucinate less with context-aware decoding. In *North American Chapter of the Association for Computational Linguistics*.

Yixuan Tang and Yi Yang. 2024. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. *Preprint*, arXiv:2401.15391.

Cunxiang Wang, Ruoxi Ning, Boqi Pan, Tonghui Wu, Qipeng Guo, Cheng Deng, Guangsheng Bao, Xiangkun Hu, Zheng Zhang, Qian Wang, and Yue Zhang. 2025. Novelqa: Benchmarking question answering on documents exceeding 200k tokens. *Preprint*, arXiv:2403.12766.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Fangyuan Xu, Kyle Lo, Luca Soldaini, Bailey Kuehl, Eunsol Choi, and David Wadden. 2024. KIWI: A dataset of knowledge-intensive writing instructions for answering research questions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12969–12990, Bangkok, Thailand. Association for Computational Linguistics.

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2024. Evaluation of retrieval-augmented generation: A survey. *ArXiv*, abs/2405.07437.

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2025. *Evaluation of Retrieval-Augmented Generation: A Survey*, page 102–120. Springer Nature Singapore.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. 2025. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM on Web Conference 2025*, pages 4442–4457.

A Implementation Details

A.1 Response Synthesis

Starting from Natural Question dataset (Kwiatkowski et al., 2019), we first filter out the instances with a passage that has more than 6,000 tokens. We then randomly select 5,500 instances from the remaining instances. For each $\alpha \in \{0, -0.5, -1, -1.4\}$, we synthesize a response according to Eq. 3, using `Qwen2.5-7B-Instruct` (Qwen, 2024). $Logit_{LLM}(* \mid \boldsymbol{q}, \boldsymbol{c})$ and $Logit_{LLM}(* \mid \boldsymbol{q})$ are modeled using in-context learning, and the in-context prompts are shown in Tab.11. Greedy search is used for sampling tokens.

Context-resistant response generation Eq. 3 can be rewritten as

$$\begin{aligned} y_i \sim softmax[&Logit_{LLM}(* \mid \boldsymbol{q}, \boldsymbol{y_{<i}}) \\ &- \beta Logit_{LLM}(* \mid \boldsymbol{q}, \boldsymbol{c}, \boldsymbol{y_{<i}}), \\ \beta = (1+\alpha)/\alpha.& \end{aligned} \tag{8}$$

We have $\alpha < -1 \Rightarrow \boldsymbol{\beta > 0}$, leading to generation that is more parametric knowledge-conditioned and context-unfaithful. On the other hand, one may question whether this holds if the model possesses the corresponding parametric knowledge of the question with high confidence, in which case both two logits (w/ and w/o $\boldsymbol{c}$) can lead to correct responses. Our method implicitly assumes that for token $\boldsymbol{y_i}$ conveying key answer information, the logits $Logit_{LLM}(\boldsymbol{y_i} \mid \boldsymbol{q}, \boldsymbol{c}, \boldsymbol{y_{<i}})$ are generally higher—or at least comparable—when supporting passages are provided than when they are absent $Logit_{LLM}(\boldsymbol{y_i} \mid \boldsymbol{q}, \boldsymbol{y_{<i}})$ (Bi et al., 2025). Therefore, for $\boldsymbol{\alpha < -1}$, even if the parametric-knowledge response is correct—for key token $\boldsymbol{y_i}$, $Logit_{LLM}(\boldsymbol{y_i} \mid \boldsymbol{q}, \boldsymbol{y_{<i}})$ is ranked top, the combined logit for key token $\boldsymbol{y_i}$ by Eq. 8 is significantly suppressed by subtracting $Logit_{LLM}(\boldsymbol{y_i} \mid \boldsymbol{q}, \boldsymbol{c}, \boldsymbol{y_{<i}})$, finally leading to outputs unfaithful to the passages $\boldsymbol{c}$.

A.2 RAG-Zeval

We utilize VERL (Sheng et al., 2024), a open-source library, to apply RL training to the models. The training runs on 8 H20 GPUs and takes approximately 20 hours. The hyperparameters for the training are listed in Tab. 6.

Hyperparameters	
Training batch size	32
Optimizer	AdamW (Loshchilov and Hutter, 2017)
Learning rate	1e-6
Warmup step	10
Gradient accumulation step	1
Learning rate scheduler	Linear
KL coefficient	0.015
Rollout temperature	1
Rollout number	8
Rollout maximum length	8192
Total epoch	2

Table 6: The settings of hyperparameters used in the RL training.

A.3 Faithfulness Metrics

When assessing the faithfulness evaluation performance of different methods, we follow Shahul et al.

(2023) to handle possible *ties* with three scenarios:

- **Best-Case**: Measures the frequency of evaluators assigning greater or equal faithfulness scores to good answers compared to poor ones.

$$best = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left[\mathrm{F}(\mathrm{good}_i) \geq \mathrm{F}(\mathrm{poor}_i)\right]$$

- **Worst-Case**: Computes the frequency of strictly greater faithfulness scores assigned to good answers.

$$worst = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left[\mathrm{F}(\mathrm{good}_i) > \mathrm{F}(\mathrm{poor}_i)\right]$$

- **Middle-Case**: Adopts ternary scoring with a partial point of 0.5 for ties.

$$middle = \frac{1}{n}\sum_{i=1}^{n}\{\mathbb{I}[\mathrm{F}(\mathrm{good_i}) > \mathrm{F}(\mathrm{poor_i})] + 0.5 \cdot \mathbb{I}[\mathrm{F}(\mathrm{good_i}) = \mathrm{F}(\mathrm{poor_i})]\}$$

A.4 Baselines

For **standard SFT** baseline, we enhance model generalizability across varying number of candidate answers by randomly partitioning the training data (described in §3.3.2) into three subsets: pairwise (2 responses), triplet (3 responses) and quadruplet (4 responses) ranking tasks. Each subset contains approximately equal data volume as reported in Tab.7. The model is trained to reproduce the relative ranking of responses based on their faithfulness with respect to the grounding passage.

	Pairwise	Triplet	Quadruplet	Total
Question	647	970	3883	5500
Instance	3877	3874	3883	11634

Table 7: Data statistics for standard supervised fine-tuning. Each original question includes four generated responses in different faithfulness levels, yielding six pairwise, four triplet, and one quadruplet ranking instance per question. The final row reports deduplicated instance counts.

For correctness evaluation, not all baseline evaluation framework has a direct correctness metric. For RAG-Checker (Ru et al., 2024), we report the performance using the *precision* metric, which aligns with our definition of correctness. For baselines without a direct correctness metric, we follow the setting in Ru et al. (2024) to report the best correlation among all metrics in Tab.2. Tab.8 shows the complete results of Trulens (Ferrara et al., 2024) and ARES (Saad-Falcon et al., 2024) using `Qwen2.5-72B-Instruct`.

Method	Model	Pearson	Spearman	Kendall
ARES	Relevancy	0.423	0.396	0.360
	Faithfulness	0.372	0.356	0.320
Trulens	Relevancy	0.368	0.320	0.289
	Faithfulness	0.428	0.446	0.360

Table 8: Complete results of Trulens and ARES with `Qwen2.5-72B-Instruct` on correctness evaluation. Performance is averaged over five trials to mitigate randomness.

B Supplemental Experiment Results

B.1 RAG-Zeval with Different Backbone Models

Besides `Qwen2.5-7B-Instruct` and `Llama-3.1-8B-Instruct` used in Tab. 1 and 2, we also apply `RAG-Zeval` (w/ and w/o RL) to different backbone models with varying parameter scales or from distinct model families, namely `Ministral-8B-Instruct-2410`, and `Qwen2.5-3B-Instruct`.

As shown in Tab. 10, across various backbone models, the RL versions of `RAG-Zeval` consistently outperform the non-RL counterparts, verifying the general efficacy of our RL-based approach. Furthermore, a comparison of the results between `Qwen-3B` and `Qwen-7B` demonstrates that the performance gains achieved by `RAG-Zeval` with RL get more pronounced when applied to larger backbone models.

B.2 Performance on Evaluation beyond Wiki

The meta evaluation dataset of correctness evaluation by Ru et al. (2024), used in our experiments, contains samples from four established datasets of various domains. The details of each split (including its domain and proportion in RAG-Checker's correctness benchmark) are as follows:

- **ClapNQ** (Rosenthal et al., 2025) (28/280) from Wikipedia;
- **KiwiQA** (Xu et al., 2024) (28/280) is specifically designed to evaluate factual consistency in retrieval-augmented generation systems within the scientific domain. The passages are sourced from peer-reviewed papers published in top-tier NLP conferences such as ACL, EMNLP, and NAACL;

Method	Backbone Model	RobustQA			KiwiQA			NovelQA		
		Pearson	*Spearman*	*Kendall*	*Pearson*	*Spearman*	*Kendall*	*Pearson*	*Spearman*	*Kendall*
BLEU	-	0.278	0.317	0.242	0.203	0.179	0.139	0.407	0.275	0.230
RougL	-	0.350	0.391	0.301	0.516	0.409	0.322	0.461	0.432	0.356
BERTScore	-	0.335	0.459	0.362	0.581	0.462	0.354	0.450	0.450	0.363
ARES	4o	0.351	0.340	0.307	0.352	0.331	0.292	0.754	0.738	0.692
	llama-70b	0.382	0.361	0.320	0.289	0.271	0.245	0.668	0.650	0.605
	qwen-72b	0.386	0.364	0.332	0.350	0.318	0.283	0.753	0.739	0.692
TruLens	4o	0.286	0.224	0.202	0.409	0.406	0.374	0.542	0.587	0.505
	llama-70b	0.307	0.211	0.192	0.278	0.273	0.251	0.750	0.679	0.618
	qwen-72b	0.341	0.278	0.252	0.303	0.298	0.273	0.584	0.559	0.516
RAGAS	embedding	0.279	0.290	0.224	0.381	0.338	0.265	0.705	0.684	0.576
RAG-Checker	4o	0.462	0.442	0.350	0.093	0.256	0.223	0.833	0.765	0.691
	llama-70b	0.399	0.358	0.280	0.260	0.346	0.275	0.750	0.772	0.678
	qwen-72b	0.453	0.431	0.342	0.260	0.304	0.260	0.880	0.840	0.751
SFT	qwen-7b	0.402	0.395	0.363	0.242	0.240	0.220	0.464	0.447	0.412
RAG-Zeval w/o RL	4o	0.553	0.533	0.433	0.386	0.391	0.305	0.853	0.846	0.782
	llama-8b	0.415	0.371	0.303	0.393	0.317	0.260	0.428	0.409	0.365
	llama-70b	0.522	0.489	0.400	0.438	0.462	0.367	0.816	0.780	0.706
	qwen-7b	0.437	0.392	0.334	0.025	-0.010	-0.010	0.704	0.657	0.593
	qwen-72b	0.541	0.520	0.432	0.575	0.581	0.472	0.839	0.796	0.717
RAG-Zeval w/ RL	llama-8b	0.547	0.480	0.378	0.539	0.440	0.341	0.562	0.528	0.442
	qwen-7b	0.535	0.500	0.391	0.277	0.247	0.193	0.770	0.721	0.721

Table 9: Performance of RAG-Zeval on evaluation datasets beyond Wikipedia

RAG-Zeval	*Model (-Instruct)*	*Best*	*Middle*	*Worst*
w/o RL	qwen-3b	0.740	0.707	0.633
	ministral-8b	0.960	0.867	0.773
w/ RL	qwen-3b	0.980	0.973	0.967
	ministral-8b	1.000	0.993	0.987

RAG-Zeval	*Model (-Instruct)*	*Pearson*	*Spearman*	*Kendall*
w/o RL	qwen-3b	0.326	0.278	0.234
	ministral-8b	0.468	0.420	0.349
w/ RL	qwen-3b	0.401	0.358	0.278
	ministral-8b	0.502	0.429	0.344

Table 10: Performance on **faithfulness** (upper) and **correctness** (lower) evaluation, assessed with different backbone models. The results are averaged over five trials to mitigate randomness.

- **RobustQA** (Han et al., 2023) (196/280) covers seven diverse domains, including Biomedical, Finance, Lifestyle, Recreation, Technology, Science, and Writing. The paper claims that none of these domains rely on Wikipedia as a source;
- **NovelQA** (Wang et al., 2025) (28/280) focuses on question answering over long-form fiction, with each question associated with a specific novel.

The last three splits present content entirely distinct from Wikipedia or encyclopedic content. We report the results for these three splits individually. Tab. 9 corroborates the generalization of `RAG-Zeval` with RL across diverse data domains, despite being trained solely on Wikipedia content.

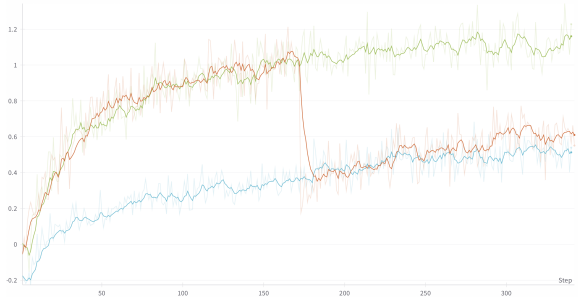

Figure 4: Reward dynamics of RL training with different data configuration. The red line represents the curriculum learning settings, while the green and blue lines are for static 3 and 4 responses, respectively.

C Training Dynamics

Figure 4 presents the reward progression during RL training under different data configurations (detailed in Tab.4). The static 4-response configuration initially yields significantly lower average rewards compared to other conditions, attributable to its greater task complexity. The curriculum-based approach (red series) experiences an expected performance dip at Step 175 during the transition to 4 candidate responses, yet maintains superior av-

erage rewards over the static 4-response baseline throughout subsequent training.

D Prompt

The prompt used in reinforcement learning is shown in Fig.5, elaborating the rules that the evaluator should conform to. Given a question, the context and K candidate answers to be assessed, the model should generate a JSON-formatted output containing detailed evaluation for each candidate answer. Each answer evaluation should involve claim decomposition, claim supportive judgment, grounding evidence generation and analysis.

Answer the following questions based on the given passages.

Question: What was the purpose of designing and building the Fiat Ecobasic concept car?

Passage: The Fiat Ecobasic is a concept car designed by the Italian manufacturer Fiat and presented in December 1999 at the Bologna Motor Show and exhibited in March 2000 at the Geneva Motor Show. The purpose of this concept was to prove that it was possible to design and build a car capable of transporting four adults in a structure made of fully recyclable composite materials and whose production and operating costs were ultra-low.

Answer: The purpose of designing and building the Fiat Ecobasic concept car was to prove that it was possible to create a car that could transport four adults using fully recyclable composite materials. Additionally, the car aimed to have ultra-low production and operating costs.

Question: When did Pope Benedict XVI become the head of the Catholic Church and sovereign of the Vatican City State, and when did he resign?

Passage: PPope Benedict XVI (Latin: Benedictus PP. XVI; Italian: Benedetto XVI; German: Benedikt XVI; born Joseph Aloisius Ratzinger; 16 April 1927 – 31 December 2022) was the head of the Catholic Church and sovereign of the Vatican City State from 19 April 2005 until his resignation on 28 February 2013. Benedict's election as pope occurred in the 2005 papal conclave that followed the death of Pope John Paul II. In 1981, he was appointed Prefect of the Congregation for the Doctrine of the Faith, one of the most important dicasteries of the Roman Curia. From 2002 until he was elected pope, he was also Dean of the College of Cardinals. Before becoming pope, he had been "a major figure on the Vatican stage for a quarter of a century"; he had had an influence "second to none when it came to setting church priorities and directions" as one of John Paul II's closest confidants.Benedict's writings were prolific and generally defended traditional Catholic doctrine, values, and liturgy. He was originally a liberal theologian but adopted conservative views after 1968. During his papacy, Benedict advocated a return to fundamental Christian values to counter the increased secularisation of many Western countries. He viewed relativism's denial of objective truth, and the denial of moral truths in particular, as the central problem of the 21st century. Benedict also revived several traditions, including the Tridentine Mass. He strengthened the relationship between the Catholic Church and art, promoted the use of Latin, and reintroduced traditional papal vestments, for which reason he was called "the pope of aesthetics". He was described as "the main intellectual force in the Church" since the mid-1980s.On 11 February 2013, Benedict announced his resignation, citing a "lack of strength of mind and body" due to his advanced age. His resignation was the first by a pope since Gregory XII in 1415, and the first on a pope's initiative since Celestine V in 1294. He was succeeded by Francis on 13 March 2013 and moved into the newly renovated Mater Ecclesiae Monastery in Vatican City for his retirement.

Answer: Pope Benedict XVI became the head of the Catholic Church and sovereign of the Vatican City State on April 19, 2005. He held this position until his resignation on February 28, 2013.

Question: `{question}`

Passage: `{Passage}`

Answer:

Answer the following questions.

Question: What was the purpose of designing and building the Fiat Ecobasic concept car?

Answer: The purpose of designing and building the Fiat Ecobasic concept car was to prove that it was possible to create a car that could transport four adults using fully recyclable composite materials. Additionally, the car aimed to have ultra-low production and operating costs.

Question: When did Pope Benedict XVI become the head of the Catholic Church and sovereign of the Vatican City State, and when did he resign?

Answer: Pope Benedict XVI became the head of the Catholic Church and sovereign of the Vatican City State on April 19, 2005. He held this position until his resignation on February 28, 2013.

Question: `{question}`

Answer:

Table 11: Prompts used to model $Logit_{LLM}(* \mid \boldsymbol{q}, \boldsymbol{c})$ and $Logit_{LLM}(* \mid \boldsymbol{q})$, respectively, for Context-Aware Decoding approach. The in-context examples are also sourced from Natural Question.

```
You are an answer evaluator for a RAG system. Given %s candidate answers, your task is to analyze
the degree to which each answer is supported by the given reference context.

### **Task:**
Conduct a strict, evidence-based evaluation of answer support using the given context.

### **Input:**
Question: <current_question>
Context: <given_context>
Answers: <K candidate_answers>

### **Evaluation Steps:**
For each answer (indexed with letters), follow this structured analysis:

1. Decompose the Answer into Atomic Claims
  - Break down the answer into one or more atomic claims—clear, standalone sentences that each
convey a single, distinct piece of information.
  - Ensure no pronouns are used; each claim must be self-contained.
  - Avoid overlap between claims; together, they should cover all necessary information without
gaps or repetition.

2. Assess Claim Support Using Context
  - For each claim, determine whether it is:
    - Supported (directly inferable from the context),
    - Unsupported (contradicted by or unrelated to the context).

3. Provide Grounding Evidence (if Applicable)
  - If a claim is supported or contradicted, extract all relevant verbatim sentences from the
context as grounding evidence that may either support or contradict it. Output the grounding
evidence as a list that is comprised of the verbatim sentences.
  - If a claim is unrelated, leave the list of grounding evidence empty.

4. Analyze Claim Support
  - For each claim, assess how clearly the context supports or contradicts it.

5. Output Valid JSON (Parsable in Python)
  - Output the analysis result in JSON format. It should be in the schema:
  ```json
  [
    {
      "id": <candidate answer id>,
      "answer": <candidate answer>,
      "atomic_claims": [
        {
          "claim": <atomic_claim>,
          "is_supported": <true/false>,
          "grounding_evidence": [
            <sentence>, ...
            ],
          "analysis": <analysis on the relation between the claim and context>
        }, ...
        ]
    },...
  ]
  ```
```

Figure 5: The complete prompt used in training the evaluator. Given the current question, context, and K candidate answers, the evaluator outputs a JSON-formatted string containing detailed evaluation for each candidate answer. Each evaluation follows the four key steps (highlighted in purple) to assess answer quality.

```
Question: price of PS3 when it first came out

Ground-truth: PlayStation3 had two hardware configurations announced: a 20 GB model and a
60 GB model , priced at US $499 ( € 499 ) and US $599 ( € 599 ).

Response A: The specific price of the PlayStation 3 when it first came out is not
mentioned in the provided content.

Response B: The text does not explicitly mention the price of PS3 when it first came out.
However, it does mention that the PS3 was initially criticized for its high price and
lack of quality games. If you're looking for the price of PS3 at launch, according to
other sources, the 20 GB model was priced at $499 and the 60 GB model was priced at $599
in the US when it was released on November 17, 2006.
```

```
[
  {
    "id": "A",
    "answer": <response A>,
    "atomic_claims": [
      {
        "claim": "The specific price of the PlayStation 3 when it first came out is not mentioned
in the provided content.",
        "is_supported": true,
        "grounding_evidence": ["PlayStation3 had two hardware configurations announced: a 20 GB
model and a 60 GB model, priced at US $499 (\u20ac499) and US $599 (\u20ac599)."],
        "analysis": "The context does not explicitly state that the specific price is not
mentioned, but it does provide the prices for the two hardware configurations. Therefore, the claim
is supported by the absence of a direct statement about the price not being mentioned."
      }
    ]
  },
  {
    "id": "B",
    "answer": <response B>,
    "atomic_claims": [
      {
        "claim": "The text does not explicitly mention the price of PS3 when it first came out.",
        "is_supported": true,
        "grounding_evidence": ["PlayStation3 had two hardware configurations announced: a 20 GB
model and a 60 GB model, priced at US $499 (\u20ac499) and US $599 (\u20ac599)."],
        "analysis": "The context does not explicitly state that the price is not mentioned, but it
does provide the prices for the two hardware configurations. Therefore, the claim is supported by
the absence of a direct statement about the price not being mentioned."
      },
      {
        "claim": "The PS3 was initially criticized for its high price and lack of quality games.",
        "is_supported": false,
        "grounding_evidence": [],
        "analysis": "The context does not mention any criticism regarding the price or the quality
of games. This claim is unsupported."
      },
      {
        "claim": "If you're looking for the price of PS3 at launch, according to other sources, the
20 GB model was priced at $499 and the 60 GB model was priced at $599 in the US when it was
released on November 17, 2006.",
        "is_supported": false,
        "grounding_evidence": [],
        "analysis": "The context does not provide this information. This claim is unsupported."
      }
    ]
  }
]
```

Figure 6: Case study (Part-1): Human annotators judge Response B as significantly better than Response A for the given question. **RAG-ZEval w/o RL** assigns response A 1 point, while response B $1/3$ point, misaligning with human preference. Additionally, the final claim produced by **RAG-ZEval w/o RL** is a direct copy of the original answer sentences, lacking atomic claim decomposition.

```
[
  {
    "id": "A",
    "answer": <response A>,
    "atomic_claims": [
      {
        "claim": "The specific price of the PlayStation 3 when it first came out is not
mentioned.",
        "is_supported": false,
        "grounding_evidence": [],
        "analysis": "The context does mention the specific prices of the PS3 models, which
contradicts this claim."
      }
    ]
  },
  {
    "id": "B",
    "answer": <response B>,
    "atomic_claims": [
      {
        "claim": "The text does not explicitly mention the price of PS3 when it first came
out.",
        "is_supported": true,
        "grounding_evidence": ["PlayStation3 had two hardware configurations announced: a
20 GB model and a 60 GB model , priced at US $499 ( \u20ac 499 ) and US $599 ( \u20ac 599
)."],
        "analysis": "The context provides the prices of the PS3 models, but it does not
explicitly state that these are the prices when it first came out."
      },
      {
        "claim": "The PS3 was initially criticized for its high price and lack of quality
games.",
        "is_supported": false,
        "grounding_evidence": [],
        "analysis": "This claim is not supported by the context and is an additional piece
of information not present in the given context."
      },
      {
        "claim": "The 20 GB model was priced at $499 in the US when it was released on
November 17, 2006.",
        "is_supported": true,
        "grounding_evidence": ["PlayStation3 had two hardware configurations announced: a
20 GB model and a 60 GB model , priced at US $499 ( \u20ac 499 ) and US $599 ( \u20ac 599
)."],
        "analysis": "The context directly supports this claim by providing the exact prices
of the PS3 models."
      },
      {
        "claim": "The 60 GB model was priced at $599  in the US when it was released on
November 17, 2006..",
        "is_supported": true,
        "grounding_evidence": ["PlayStation3 had two hardware configurations announced: a
20 GB model and a 60 GB model , priced at US $499 ( \u20ac 499 ) and US $599 ( \u20ac 599
)."],
        "analysis": "The context directly supports this claim by providing the exact prices
of the PS3 models."
      }
    ]
  }
]
```


Figure 7: Case study (Part-2): For the same question and responses (shown in Fig. 6), **RAG-Zeval** assigns response A 0 point and response B $3/4$ point, producing a ranking consistent with human judgment.