

# YESciEval: Robust LLM-as-a-Judge for Scientific Question Answering

Anonymous ACL submission

## Abstract

Large Language Models (LLMs) drive scientific question-answering on modern search engines, yet their evaluation robustness remains underexplored. We introduce **YESciEval**, an open-source framework that combines fine-grained rubric-based assessment with reinforcement learning to mitigate optimism bias in LLM evaluators. We release multidisciplinary scienceQ&A datasets, including adversarial variants, with evaluation scores from multiple LLMs. Independent of proprietary models and human feedback, our approach enables scalable, cost-free evaluation. By advancing reliable LLM-as-a-judge models, this work supports AI alignment and fosters robust, transparent evaluation essential for scientific inquiry and artificial general intelligence.

## 1 Introduction

The rise of scientific search engines powered by generative Large Language Models (LLMs)—such as **Elicit**, **OpenScholar**, **SciSpace**, and **ORKG Ask**—has transformed how researchers search and synthesize scholarly information. A key feature of these platforms is scientific question answering (scienceQ&A), where an LLM synthesizes insights from top-ranked papers to generate concise responses (Pride et al., 2023; Giglou et al., 2024). While aligning LLMs to human values (e.g., helpfulness, harmlessness, honesty) is well studied (Askell et al., 2021; Zheng et al., 2023), their real-world robustness in scienceQ&A remains largely unexplored. The domain-agnostic and free-form nature of scienceQ&A limits the applicability of traditional n-gram-based metrics (e.g., BLEU (2002a), ROUGE (2004a)), which may not fully capture compositional and domain-specific reasoning (Krishna et al., 2021). Human evaluation, though more nuanced, is costly and difficult to scale (Krishna et al., 2023). Recent efforts using LLMs as evaluators (LLM-as-a-judge (Zheng et al., 2023))

show parity with human judgment (Chiang and Lee, 2023b) but also exhibit biases (Gudibande et al., 2023; Ye et al., 2024), highlighting the need for a reliable and fair LLM-based evaluation system. Most prior evaluation work relies on proprietary GPT models (Wang et al., 2023a; Dubois et al., 2023; Liu et al., 2023; Fu et al., 2024), raising issues of transparency and high costs at scale (Kim et al., 2023, 2024). To address these challenges, we propose a framework that pairs fine-grained rubrics with an open-source LLM-as-a-judge methodology, enabling more reliable and cost-effective evaluation of scienceQ&A across diverse scientific domains.

Despite the growing use of LLMs for scienceQ&A, a systematic evaluation framework remains absent, limiting iterative development and transparency. Such a framework would provide feedback for reward-based training and help users assess the reliability of automated answers, mitigating misinformation. While human evaluations are the gold standard, they are costly and hard to scale (Ouyang et al., 2022; Krishna et al., 2023), underscoring the need for an automated, trustworthy alternative. An LLM-as-a-judge approach with fine-grained rubrics can address this gap, enabling robust, repeatable assessments that enhance scienceQ&A systems and user confidence.

To establish a systematic and transparent evaluation framework for scienceQ&A, we propose **YESciEval**, which integrates two core components. First, we define a **nine-rubric multifaceted assessment scheme**, structured into three key dimensions—*Linguistic & Stylistic Quality*, *Logical & Structural Integrity*, and *Content Accuracy & Informativeness*—to comprehensively evaluate LLM-generated responses. Each rubric is scored on a Likert scale (1–5) with predefined guidelines, ensuring consistent and rigorous assessment. Second, we address the **optimism bias** (Villafior et al., 2022) inherent in LLMs, which might be counterproductive to LLM-as-a-judge role, where models,

designed as conversational agents, favor positive engagement over critical assessment. The title prefix of this paper, **YESci** (pronounced ‘yes, sigh!’), playfully encapsulates our central research question: *how can we mitigate LLMs’ optimism bias and enhance their robustness as scienceQ&A evaluators?* To tackle this, we employ a two-step alignment strategy: **supervised fine-tuning followed by reinforcement learning (RLHF)**, ensuring adherence to our fine-grained rubrics for reliable evaluation. Additionally, we introduce **adversarial datasets**—systematically constructed instances where LLMs are expected to underperform—to *contrast standard and adversarial responses*, reinforcing robustness against evaluation errors. While our approach is applied to scienceQ&A in this study, mitigating optimism bias in LLM evaluators has broader implications for other evaluation tasks. By integrating *fine-tuning, reinforcement learning, and adversarial perturbations*, **YESciEval** enhances LLM-based evaluation reliability, particularly for open-source models, reducing reliance on proprietary systems and providing a scalable, transparent framework for scientific inquiry.

We select four open-source LLMs spanning 8B to 123B parameters (e.g., LLaMA-3.1 and Mistral-Large) from Meta AI, Mistral AI, and Alibaba Cloud for generating (**LLM<sub>gen</sub>**) and evaluating (**LLM<sub>eval</sub>**) scienceQ&A content. Each model in the generator role produces a unique scienceQ&A dataset with its responses as the benign (non-perturbed) dataset, and we introduce two adversarial perturbation types—**extreme** (overt distortions) and **subtle** (lightweight heuristics)—, yielding 12 datasets. Rotating the four models as evaluators and scoring responses under our nine-rubric framework results in 48 LLM-as-a-judge configurations. To validate generalizability, we apply a two-step alignment strategy (*supervised fine-tuning* followed by *reinforcement learning*) to LLaMA 3.1 8B as an evaluator, confirming our approach’s robustness for different model sizes and iterations. The release of **LLaMA 3.3** underscores the importance of flexible, model-agnostic frameworks like YESciEval.

Despite significant advances in the generative capabilities of LLMs, our findings reveal that they remain unexpectedly fragile when confronted with basic, heuristic-based adversarial attacks in the role of a judge. Against this backdrop of LLM-as-a-judge for scienceQ&A, we pose the following research questions. **RQ1: How similar are scienceQ&A responses across different LLM**

**families?** Given the rapid influx of new LLMs, we aim to reduce the uncertainty around model choice by clarifying the degree of similarity in how different model families handle scienceQ&A. **RQ2: How do LLM-as-judge evaluations correlate for the benign scienceQ&A setting?** Beyond examining alignment in evaluative behavior, we investigate whether LLMs exhibit bias toward their own generated answers. **RQ3: Can a smaller open-source LLM be adapted as an LLM-as-a-judge for scienceQ&A while overcoming optimism bias?** While prior work focuses on tuning LLMs to specific skills (rubrics), a key challenge remains curtailing their *optimism bias*. We propose a framework that integrates supervised fine-tuning, reinforcement learning, and adversarial alignment to equip smaller open-source models with robust evaluative capabilities. Although we illustrate our approach in the context of scienceQ&A, these methods can be adapted to other generative AI scenarios requiring rigorous critique.

This work makes the following key contributions: 1. **Multidisciplinary benign and adversarial scienceQ&A datasets** – We release a comprehensive scienceQ&A corpus with adversarial variants to evaluate LLM robustness, available on [GitHub](#). 2. **Comprehensive evaluation benchmark** – We provide evaluation scores and rationales from multiple LLMs in both vanilla and adversarial settings, supporting further research and reproducibility. 3. **Optimism bias mitigation** – We implement a reinforcement learning framework to align LLM evaluation behavior with real-world critical feedback expectations. The YESciEval source code is available on [GitHub](#). 4. **Scalable, cost-free evaluation paradigm** – Our approach is independent of proprietary models and human feedback, leveraging open-source LLMs on nationwide supercomputing infrastructure. While human feedback is invaluable for LLM alignment, it is often infeasible to obtain. We present a zero-cost alternative, integrating rubric-based evaluation with adversarial testing to ensure reliable LLM-as-a-judge models. This eliminates experimental costs, aside from researcher time and compute resources for running open-source LLMs.

This research presents a reproducible, cost-free framework for evaluating natural language generation (NLG) in scienceQ&A, advancing AI alignment, robustness, and the broader agenda of LLM plausibility while pushing the frontiers of artificial general intelligence (AGI).

## 2 Task Definition

The YESciEval framework for scienceQ&A consists of two tasks: *LLMgen* for generating responses and *LLMeval* for evaluating them.

### Task 1: ScienceQ&A Generation (*LLMgen*)

Generates a synthesized summary response  $A$  to a research question  $Q$  using abstracts from the top  $N$  relevant papers. It must demonstrate (1) domain knowledge, (2) numerical proficiency, (3) long-range context understanding, and (4) cause-and-effect reasoning (Wadden et al., 2020).

### Task 2: *LLMeval* (ScienceQ&A Evaluation)

Assesses the quality of  $A$  based on predefined rubrics and context as  $Q$  and  $N$  abstracts.

The evaluation prompt is structured as follows: (1) *Context*, defining scienceQ&A synthesis as the generation of a coherent summary from research papers to address  $Q$ ; (2) *Role*, assigning *LLMeval* as the evaluator; (3) *Task Description*, ensuring  $A$  accurately synthesizes information from abstracts; (4) *Evaluation Characteristics*, specifying the rubric applied; (5) *Rating Scale*, using a 1–5 Likert scale with rubric-specific guidelines; (6) *Response Format*, requiring structured ratings and rationales in JSON format; and (7) *Notes*, emphasizing objectivity and adherence to source content.

The output consists of a score  $S$  (1–5) and a rationale.

## 3 The YESciEval Qualitative Rubrics

Drawing from a comprehensive review of evaluation rubrics in prior LLMs-as-a-judge research (see Related Work subsection 7.2), we define a nine-rubric qualitative evaluation paradigm for YESciEval. Chosen for simplicity, memorability, and precise definability, these rubrics minimize overlap while capturing key facets of an ideal LLM response to science Q&A. Each is framed as a concise question to reduce ambiguity for LLM-as-a-judge or human evaluators. Organized into three main evaluation dimensions, they are as follows.

**Linguistic & Stylistic Quality** concerns grammar, clarity, and adherence to academic writing conventions. This category comprises three rubrics: 1. *Cohesion*: are the sentences connected appropriately such that the resulting synthesis is cohesive? 2. *Conciseness*: is the answer short and clear, without redundant statements? 3. *Readability*: does the answer follow appropriate style and structure conventions for academic writing, particularly for read-

ability? **Logical & Structural Integrity** focuses on the reasoning and organization of information. This category comprises three rubrics: 4. *Coherence*: are the ideas connected in a sound and logical manner? 5. *Integration*: are the sources structurally and linguistically well-integrated, using appropriate markers of provenance/quotation and logical connectors for each reference? In addition, are the sources integrated as a single paragraph? 6. *Relevancy*: is the information in the answer relevant to the question? **Content Accuracy & Informativeness** ensures that the response is both correct and useful. This category comprises three rubrics: 7. *Correctness*: is the information in the answer a correct representation of the content of the provided abstracts? 8. *Completeness*: is the answer a comprehensive encapsulation of the relevant information in the provided abstracts? 9. *Informativeness*: is the answer a useful and informative reply to the question?

For each rubric, the LLM-as-a-judge rates response quality on a Likert scale from 1 (very bad) to 5 (very good), with predefined guidelines ensuring consistency. These guidelines set clear expectations for each level—for instance, in *Readability*, a 1 indicates severe issues in style, structure, and language use, while a 5 reflects an exceptionally well-written, academically sound synthesis. Standardized criteria help both LLMs and human evaluators apply consistent judgment. Full rating guidelines for all nine rubrics are provided in Appendix E. Ideally, an LLM-as-a-judge assigns perfect scores across all rubrics, while suboptimal responses receive lower ratings based on specific deficiencies.

## 4 Science Q&A Datasets

After reviewing existing Q&A datasets (see Related Work subsection 7.1), we selected two that meet our scienceQ&A task definition: the ORKGSynthesis dataset (2024) (ORKGSyn) and the BioASQ dataset (2024a). ORKGSyn offers a diverse set of crowdsourced research questions for evaluating LLMs’ scienceQ&A capabilities, while BioASQ provides handcrafted QA pairs spanning four question types. Both were chosen because they include science questions  $Q$  with each question linked to  $N$  relevant papers.

### 4.1 The ORKGSynthesis Corpus

This corpus comes from the domain-expert-curated structured research comparisons (Oelen et al.,



Research field	Frequency
Computer Sciences	125
Physics	28
Animal Sciences	19
Chemistry	17
Urban Studies and Planning	16
Earth Sciences	14
Oceanography and Atmospheric Sciences and Meteorology	14
Science and Technology Studies	12
Materials Science and Engineering	12
Engineering	10

Table 1: Top 10 domains in the ORKGSynthesis dataset.

2019) on the Open Research Knowledge Graph (ORKG) platform (Auer et al., 2020). Its accompanying LLM-powered search engine, ORKG Ask, synthesizes abstracts from the top five papers for any given research question. Building on this approach and the ORKG as a gold-standard source, prior work (Giglou et al., 2024) compiled a dataset of 348 entries—each linking a research question with exactly five relevant papers. Because the ORKG spans multiple disciplines, the dataset covers 33 research fields. Table 1 lists the top 10 fields; the full list appears in Figure 4 in subsection B.1.

While ORKGSyn is multidisciplinary, the next corpus is in biomedicine.

## 4.2 The BioASQ Corpus

BioASQ (Nentidis et al., 2024a) is an annual biomedical semantic indexing and Q&A challenge. Its 2024 edition covers four NLP tasks, and we focus on the first: biomedical Q&A. The dataset includes 5389 domain-expert-curated Q-A pairs in four question types: “yes/no,” “factoid,” “list,” and “summary.” The challenge has three phases (A, A+, and B), with Phase B linking each question to a human-annotated set of relevant papers (Nentidis et al., 2024b), meeting our requirement of  $Q$  with  $N$  relevant papers. Because  $N$  varies in BioASQ, we capped it at 40 for computational feasibility and to fit LLM input context (see Figure 5, subsection B.2). This variability introduces a unique scienceQ&A setting compared to ORKGSyn. We narrowed the dataset to the test set’s 73 “summary” questions, each with up to 40 PubMed abstracts.

## 4.3 Our ScienceQ&A Dataset Compilation

Now that we had corpora of  $Q$  linked with  $N$  paper abstracts, we need to apply  $LLM_{gen}$ .

**$LLM_{gen}$  models.** Our selection criteria for LLMs were simple: they had to be open-source, state-of-the-art at some point, and diverse in size,

including at least one small model. Based on this, we chose Llama 3.1 8B & 70B (MetaAI, 2024), Qwen 2.5 72B (AlibabaCloud, 2024), and Mistral Large 128B (MistralAI, 2024). All feature 128K-token context windows and excel in reasoning, coding, and multilingual tasks. Llama 3.1 prioritizes efficiency and safety with Llama Guard 3, Qwen 2.5 offers robust multilingual support (29+ languages) and a specialized Coder variant, while Mistral Large 128B supports 80+ programming languages and is optimized for single-node inference. While Llama 3.1 emphasizes open-source accessibility, Qwen 2.5 specializes in multilingual and coding capabilities, and Mistral Large delivers peak computational performance despite its size.

**$LLM_{gen}$  task.** The four models were applied, in turn, to generate a synthesized summary response,  $A$ , for a given research question,  $Q$ , using abstracts from the top  $N$  relevant papers. This process was conducted for both ORKGSyn and BioASQ. The resulting datasets, released in the dataset folder of our code repository, constitute the “benign” science Q&A dataset variant of this study.

## 4.4 Our Adversarial ScienceQ&A Corpus

Adversarial attacks in NLP generate text samples designed to mislead models into incorrect predictions, using either heuristic-based perturbations or machine learning algorithms. Inspired by Jia and Liang (2017), who applied heuristic adversarial perturbations to assess NLP brittleness in reading comprehension on SQuAD (2016), we introduce rubric-based adversarial variants of the benign scienceQ&A dataset to evaluate the robustness of LLM-as-a-judge systems.

A key question in this study is the reliability of  $LLM_{eval}$ ’s qualitative scores for benign synthesized answers  $A$  generated by  $LLM_{gen}$ . Traditionally, human evaluations would serve as a benchmark, but instead, we assess LLM-as-a-judge robustness through adversarial testing. This approach indirectly measures reliability by evaluating whether  $LLM_{eval}$  appropriately differentiates between benign and perturbed responses. If the model fails to adjust its scores accordingly, it suggests an inability to critique responses effectively, thus low reliability on the LLM-as-a-judge.

To systematically evaluate this, we adopt a heuristic-based perturbation strategy, introducing deliberate errors at specific points in the benign text with the expectation that evaluation scores should reflect the resulting quality deterioration. To en-

sure rigor in adversarial assessment, we design two variants of adversarial attacks: (1) *Subtle adversarial samples*, where minor alterations to the benign text create realistic yet difficult-to-detect errors, and (2) *Extreme adversarial samples*, where substantial modifications make flaws obvious. The adversarial perturbation heuristics, for both variants, were designed at the fine-grained rubric level. Thus each of the nine rubrics have *subtle* and *extreme* adversarial perturbation heuristics associated with them. They are as follows. **Relevancy** assesses whether the response remains on-topic, with subtle attacks appending sentences from related synthesis paragraphs and extreme attacks injecting unrelated sports news. **Correctness** tests factual alignment with provided abstracts, using the same attack strategy. **Completeness** measures how well the response encapsulates relevant content, with subtle attacks removing the final sentence and extreme attacks also appending unrelated text. **Informativeness** evaluates the response’s utility, using the same perturbation as relevancy. **Integration** examines structural and linguistic coherence, with subtle attacks removing the first logical connector and extreme attacks eliminating all connectors. **Cohesion** ensures appropriate sentence connections, with subtle attacks swapping the last two sentences and extreme attacks randomly shuffling them. **Coherence** assesses logical idea progression, using the same attack as relevancy. **Readability** focuses on adherence to academic writing conventions, with subtle attacks adding informal blog snippets and extreme attacks inserting tweets. **Conciseness** tests redundancy, with subtle attacks appending an LLM-generated redundant version of the last sentence and extreme attacks inserting redundant text after every sentence. Detailed heuristics are provided in Appendix subsection B.3. These perturbations systematically degrade response quality, ensuring rigorous evaluation of LLM-as-a-judge reliability. Redundant responses were LLM-generated and manually refined for fluency. The adversarial dataset, [publicly available](#), benchmarks LLMs’ ability to mitigate optimism bias and distinguish response quality. Unlike synonym substitution attacks (SSAs) (Alzantot et al., 2018), which are conjectured to degrade fluency and meaning (Chiang and Lee, 2023a), our rubric-based heuristics introduce syntactic violations to ensure low-quality text. This controlled degradation enables precise evaluation of LLM score adjustments in scienceQ&A assessment.

	BioASQ	ORKGSynthesis
$LLM_{gen}$ Train	51	234
$LLM_{gen}$ Test	22	105
$LLM_{eval}$ Train Sets		
SFT	6,504	34,991
RL (adversarial)	1,669	6,148
RL (benign+adversarial)	2,569	2,290
$LLM_{eval}$ Test Set	2,376	11,340

Table 2: Dataset statistics across training and test sets for  $LLM_{gen}$  and  $LLM_{eval}$ .

#### 4.5 Our LLM-as-a-judge Evaluations

The four models—Llama 3.1 8B & 70B, Qwen 2.5 72B, and Mistral Large 128B—previously used for  $LLM_{gen}$  were now tasked with  $LLM_{eval}$ . section 2 details the task specification prompt. Each model evaluated all instances from ORKGSyn (benign, subtle-adv, extreme-adv) and BioASQ across the nine YESciEval rubrics. This resulted in 37,584 evaluation scores for ORKGSyn ( $348 \times 3 \times 9 \times 4$ ) and 7,884 for BioASQ ( $73 \times 3 \times 9 \times 4$ ), equating to 9,396 and 1,971 evaluations per model, respectively. Notably, despite requiring GPU compute, model access incurred zero monetary cost—whereas using proprietary models like OpenAI’s GPT would have cost at least 1,000 euros or dollars for these evaluations. This reinforces the motivation of YESciEval, aligned with prior studies (Kim et al., 2023, 2024), to enhance open-source LLM-as-a-judge models for greater accessibility.

### 5 The YESciEval Alignment Method

This work aims to propose a cost-free evaluation framework without human annotators or proprietary models. Key contributions include an adversarial testing strategy for robust LLM-as-a-judge models and the YESciEval alignment method discussed in this section that can be applied to open-source LLMs to equip them as robust evaluators.

The methodology is remotely related to self-instruct (Wang et al., 2023b), where LLMs are improved for instruction following self-generated instructions; except we reinforce evaluation behaviors of LLMs to desired behaviors by treating undesired behaviors as negative reward signals. The Figure 1 depicts our RL technique applied to learn the alignments between desirable and undesirable behaviors with the following steps:

**Supervised Fine-Tuning (SFT).** As baseline and starting point for the evaluator model, we fine-tuned *LLaMA-3.1-8B* model with structured data (stats presented in Table 2) from  $x : LLM_{gen}$  and

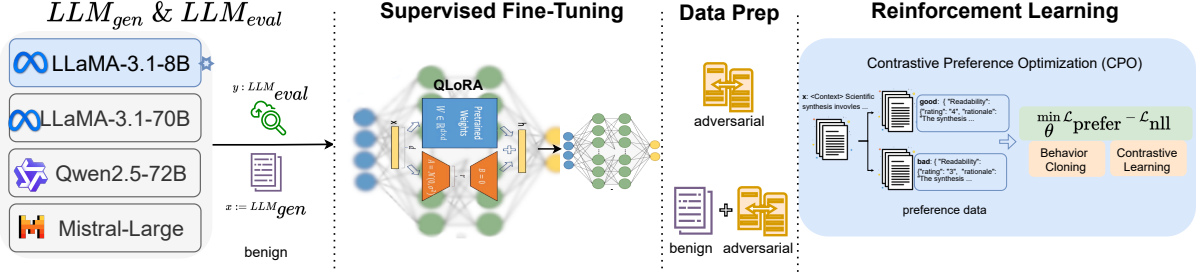


Figure 1: YESciEval LLM-as-a-Judge Alignment: Supervised fine-tuning of  $LLM_{eval}$ , followed by reinforcement learning via Contrastive Preference Optimization to align open-source LLMs with desired rubric-level evaluations.

$y : LLM_{eval}$  for simulated preference alignments. Here  $x : LLM_{gen}$  is the benign scientific synthesis from four LLMs with respective evaluation from  $LLM_{eval}$  for nine rubrics which  $y$  considers each evaluation rubric as a separate sample for SFT. The fine-tuning objective at this stage is supervised learning, where the model is trained to efficiently adapt large models while keeping computational costs manageable using Quantized Low-Rank Adaptation (QLoRA) (Dettmers et al., 2023).

**Data Preparation for RL.** Unlike traditional fine-tuning, RL does not require large datasets; instead, models learn from comparisons rather than absolute labels, reducing redundancy and cost (Ziegler et al., 2019). To this end, the data is divided into benign and adversarial samples, ensuring that our models reinforce the distinction between 'good' (desirable) and 'bad' (undesirable) evaluations as it is critical for preference modeling (Askell et al., 2021). We impose a 100-per-rubrics, per- $LLM_{eval}$  threshold (for ORKGSyn in RL (adversarial) this threshold is set to 500 due to the large nature of the task) to maintain a manageable dataset size while preserving diversity across criteria. The obtained dataset stats are represented in Table 2 for two different sets, where in adversarial sets, for the extreme adversarial dataset we set desirable evaluation score for nine-rubrics as a 1, while for the subtle adversarial dataset, this threshold is set to  $\leq 3$ . This would create a clear distinction between good and bad evaluations while accounting for varying degrees of evaluation quality. So, for any evaluations from  $LLM_{eval}$ , if a rating is above the specified threshold, it will be considered as a bad evaluation sample otherwise, it will be used as a good evaluation sample. However, RL can mimic only the bad examples and avoid imitating the benign synthesis evaluation, where there are no adversarial settings applied, so an-

other set with benign and adversarial is curated, whereas bad samples are selected based on undesirable evaluations from adversarial. The final dataset of  $\mathcal{D} = \{x_{LLM_{gen}}^{(i)}, y_{good}^{(i)}, y_{bad}^{(i)}\}_{i=1}^N$  is constructed for RL technique.

**Reinforcement Learning.** Beyond SFT, the RL technique is required to further align LLM's evaluations with desirable/undesirable behaviors instead of SFT's reference-mimicking learning process. Here, the key idea is to calibrate evaluation capabilities by modeling preference data  $\mathcal{D}$ . In this step, Imitation Learning (IL) is applied through Contrastive Preference Optimization (CPO) (Xu et al., 2024) to effectively learn alignment from  $\mathcal{D}$ . CPO is a state-of-the-art method that enhances Direct Preference Optimization (DPO) (Rafailov et al., 2023) by teaching the model to identify and reject suboptimal examples. Unlike traditional methods that focus solely on generating preferred responses, CPO uses contrastive learning with hard negative examples to help the model distinguish between high-quality and low-quality outputs. This approach improves the model's ability to prioritize superior responses while effectively rejecting lesser ones, leading to more discriminative evaluations. Considering  $\pi_{\theta}$  as a parameterized policy, the CPO loss is defined as  $\min_{\theta} \underbrace{\mathcal{L}(\pi_{\theta}, U)}_{\mathcal{L}_{prefer}} - \underbrace{\mathbb{E}_{(x, y_{good}) \sim D} [\log \pi_{\theta}(y_{good}|x)]}_{\mathcal{L}_{NLL}}$ ,

where  $\mathcal{L}_{prefer}$  is a behavior cloning (BC) (Hejna et al., 2023) regularization that encourages the policy  $\pi_{\theta}$  to align with uniform prior distribution  $U$ , derived from reference policy  $\pi_{LLM_{eval}}$  by preferring good evaluation  $y_{good}$  over bad one  $y_{bad}$ . Moreover, the term  $\mathcal{L}_{NLL}$  is the negative log-likelihood loss that penalizes the policy for making poor generation of the action  $y_{good}$  from given state  $x$  for high-quality judgments.



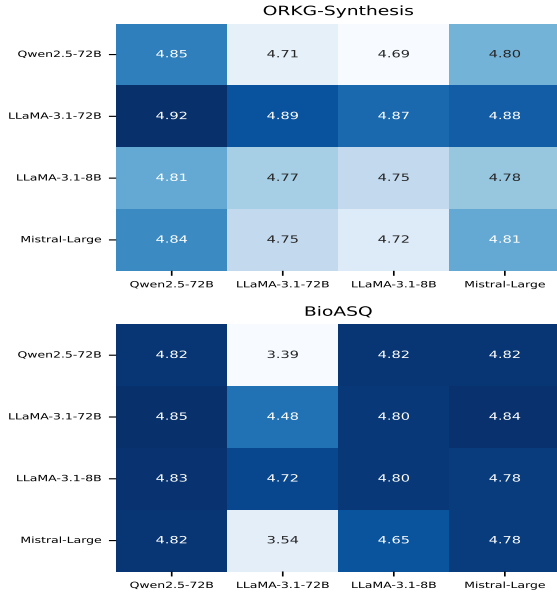


Figure 2: Heatmaps depicting agreement for synthesis evaluations on benign datasets. The x-axis represents the  $LLM_{gen}$ , while the y-axis denotes the  $LLM_{eval}$ .

## 6 Results and Discussion

In this section, we systematically analyze the results in relation to the three main research questions outlined in the Introduction. Specifically, we discuss in detail observations on the results obtained from the two-stage process:  $LLM_{gen}$  (RQ1) and  $LLM_{eval}$  (RQ2); and the application of our YESciEval LLM-as-a-Judge alignment method (RQ3). For details on our experimental setup and training, we refer the reader to Appen. F and G.

First, we focus on: **RQ1: How similar are scienceQ&A responses across the three different LLM families when applied as  $LLM_{syn}$ ?** To address this RQ, we measured similarities between benign syntheses generated by the four models for ORKGSyn and BioASQ separately. Based on a comprehensive review of NLG metrics (Sai et al., 2022), we applied eight diverse similarity metrics: four for verbatim matching, one edit-distance-based, and three embedding-based, e.g., MoverScore (2019), BERTScore (2020). To visualize  $LLM_{syn}$  correlations per dataset, we computed confusion matrices with averaged similarity scores.

Overall, ORKGSyn consistently yielded higher alignment scores than BioASQ, likely due to its broader domain coverage, with Computer Science (125 questions) as the largest category. As AI and digital fields grow, general-purpose LLMs trained on large datasets, including research papers, exhibit

stronger proficiency in these areas. In contrast, BioASQ’s biomedical focus, a more specialized domain, led to greater uncertainty and lower correlation scores. Strong correlations emerged within model pairs: Llama 8B and 70B, likely due to their shared Meta origin and training data, differing mainly in parameter size, and Mistral and Qwen, suggesting overlapping training data. These findings highlight the role of shared training data and architecture in aligning model outputs across LLM families. Detailed results: 16 confusion matrices across eight metrics are in Appendix D.

**RQ2: How do LLM-as-judge evaluations correlate for the benign scienceQ&A setting?** To address this question, we analyze the results in Figure 2, focusing on  $LLM_{eval}$  outputs. Each LLM evaluated the benign synthesis dataset created by the four models in their  $LLM_{syn}$  roles, with results presented as a confusion matrix where each cell represents the averaged score across all rubrics and synthesis instances, mapping evaluator LLMs (y-axis) against  $LLM_{syn}$  models (x-axis). Across both datasets, evaluators assigned higher scores to BioASQ than ORKGSyn, likely due to ORKGSyn’s interdisciplinary nature, where  $LLM_{syn}$  models struggled with certain domains. Notably, no  $LLM_{eval}$  exhibited bias toward its own generated synthesis (Li et al., 2025). Instead, all evaluators consistently preferred synthesis outputs generated by Qwen. As the largest model, Qwen likely demonstrates superior text generation abilities, reinforcing the reliability of LLM-as-a-judge, even for smaller models like Llama 8B.

**RQ3: Can a smaller open-source LLM be adapted as an LLM-as-a-judge for scienceQ&A while overcoming optimism bias?** To address this question, we present a comprehensive results plot in Figure 3, where each row depicts  $LLM_{eval}$  results across three dataset variants, aggregated per rubric for BioASQ (top row) and ORKGSyn (bottom row). The goal is to assess the efficacy of the YESciEval alignment method on a small LLM, specifically Llama 8B, under the premise that if effective on a lower-parameter model, it should generalize to larger LLMs. The highlighted colored lines represent different Llama 8B variants: blue (vanilla model), yellow (SFT model trained on benign data), red (SFT + RL with adversarial alignment), and green (SFT + RL with a balanced subsample of benign and adversarial evaluations). Light gray lines indicate the vanilla model performances of the other three LLMs. As hypothesized,

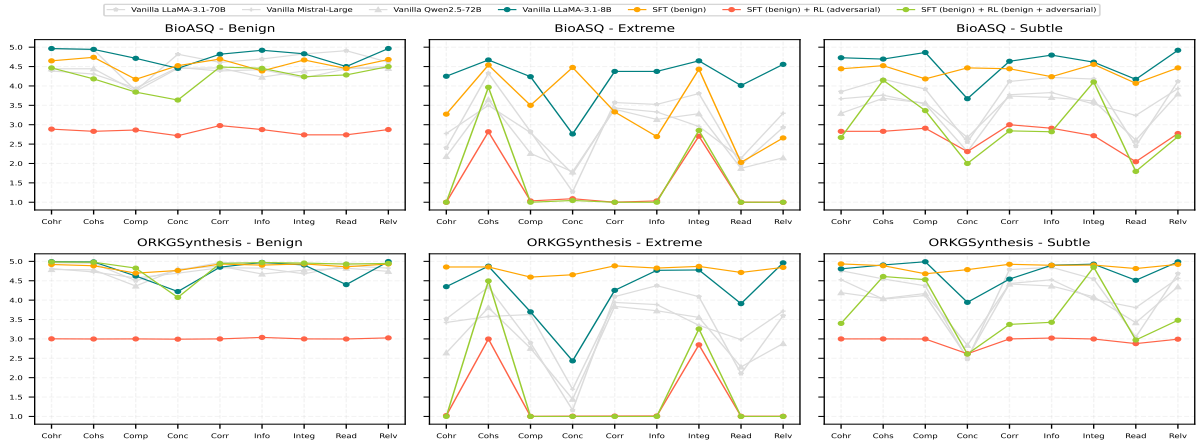


Figure 3: Evaluation of synthesis across different models and fine-tuning strategies on BioASQ and ORKGSynthesis datasets. The nine rubrics include Coherence (Cohr), Cohesion (Cohs), Completeness (Comp), Conciseness (Conc), Correctness (Corr), Informativeness (Info), Integration (Integ), Readability (Read), and Relevancy (Relv).

vanilla Llama 8B (blue) exhibited excessive optimism, assigning high scores even in extreme adversarial cases—e.g., scoring above 4 on Corr (correctness rubric) despite perturbations introducing unrelated sports news sentences. Fine-tuning on benign data alone (yellow) further amplified optimism, necessitating alignment. When RL was applied only to adversarial data (red), the model became overly pessimistic. However, when RL was trained on both benign and adversarial samples (green), Llama 8B stabilized as a robust evaluator, addressing the RQ. It assigned relatively high scores for benign syntheses while distinguishing adversarial perturbations, scoring around 1 in extreme cases and around 3 in subtle cases, demonstrating rubric-specific discrimination.

## 7 Related Work

### 7.1 Question & Answering Benchmarks

Automatic Q&A spans diverse datasets varying in domain and Q&A type. Of 41 NLP Q&A datasets reviewed by Wang (2022), only BioASQ aligns with scienceQ&A. Multiple-choice (e.g., PubMedQA (2019), MMLU (2021)), Boolean (e.g., BoolQ (2019)), and numerics (e.g., Math Dataset (2019)) fall outside our scope, as do bibliographic (Banerjee et al., 2023) and knowledge graph extraction datasets (Auer et al., 2023; Yan et al., 2024). Existing benchmarks, such as Hugging Face leaderboard tasks (Wang et al., 2024; Rein et al., 2023) and alignment-focused chat-based evaluations (Reddy et al., 2019; Zheng et al., 2023; Köpf et al., 2024), primarily assess multiple-choice reasoning or human preference alignment. In contrast,

we introduce a generative scienceQ&A dataset, filling a gap in current benchmarking efforts.

### 7.2 LLM-as-a-judge Evaluation Rubrics

LLM-as-a-judge (Zheng et al., 2023) initially focused on correlating LLM evaluations with human judgments in open-domain NLG, primarily using pairwise preference evaluations (Wang et al., 2023a; Chiang and Lee, 2023b; Dubois et al., 2023; Liu et al., 2023). Some works incorporated rubrics, such as G-Eval (Liu et al., 2023) for summarization and GPTScore (Fu et al., 2024), which aligns closely with our criteria. Recent frameworks emphasize fine-grained rubrics; FLASK (Ye et al., 2024) assesses robustness, correctness, efficiency, factuality, and readability, of which nine align with our work. Prometheus (Kim et al., 2023, 2024) expands rubric-based evaluation but relies on human references, whereas we use adversarial data to refine LLM evaluations without annotations. ScienceQ&A assessment evolved from three core criteria—comprehensiveness, trust, and utility (Pride et al., 2023)—to the nine rubrics adopted in this work (Giglou et al., 2024). Decoupling reliance on human references, our rubric-based adversarial approach provides a systematic, cost-free framework for scienceQ&A evaluation.

## 8 Conclusion

The YESciEval approach is a demonstrated reproducible, cost-free framework for evaluating NLG in scienceQ&A, advancing AI alignment, robustness, and the broader agenda of LLM plausibility while pushing the frontiers of AGI.



## 9 Limitations

While our approach establishes a robust, cost-free framework for LLM-as-a-judge evaluation in scienceQ&A, certain methodological choices present limitations and opportunities for future improvement. One avenue worth exploring is **chain-of-thought (CoT) reasoning** (Wei et al., 2022), which has demonstrated efficacy in structured reasoning tasks, such as mathematics and logic problems (e.g., SelfCheck (Miao et al., 2024)). While CoT is not yet widely applied to evaluation tasks, it may enhance LLM judgment by promoting more structured decision-making. We plan to investigate CoT-enhanced evaluators in future work.

Another consideration is our choice of reinforcement learning (RL) over *few-shot in-context learning* (ICL) (Brown et al., 2020) for improving evaluation robustness. While ICL, where models are conditioned on good and bad examples (Fu et al., 2024), is a promising technique, it poses practical challenges in our setup. Given our **nine rubrics**, an ICL approach would require significantly longer prompt contexts, potentially exceeding the model’s ability to effectively utilize the information. Additionally, curating high-quality bad examples would necessitate human annotations, contradicting our goal of maintaining a fully automated, zero-cost alignment process. Despite these concerns, we acknowledge the potential of ICL and plan to explore more efficient adaptations in future iterations.

Finally, while our approach reduces reliance on proprietary models, it remains constrained by the capabilities of current open-source LLMs, which may not fully match the performance of commercial alternatives. As stronger open-source models emerge, refining evaluator alignment strategies will be critical to ensuring continued robustness and reliability in LLM-based evaluation. Future work will also investigate generalizing our methodology beyond scienceQ&A to broader generative AI evaluation tasks.

## References

- AlibabaCloud. 2024. Qwen2.5: A party of foundation models! <https://qwenlm.github.io/blog/qwen2.5/>. Accessed: 16 February 2025.
- Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on*

*Empirical Methods in Natural Language Processing*, pages 2890–2896.

- Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.
- Sören Auer, Dante AC Barone, Cassiano Bartz, Eduardo G Cortes, Mohamad Yaser Jaradeh, Oliver Karras, Manolis Koubarakis, Dmitry Mouromtsev, Dmitrii Plukhin, Daniil Radyush, et al. 2023. The sciqa scientific question answering benchmark for scholarly knowledge. *Scientific Reports*, 13(1):7240.
- Sören Auer, Allard Oelen, Muhammad Haris, Markus Stocker, Jennifer D’Souza, Kheir Eddine Farfar, Lars Vogt, Manuel Prinz, Vitalis Wiens, and Mohamad Yaser Jaradeh. 2020. Improving access to scientific literature with knowledge graphs. *Bibliothek Forschung und Praxis*, 44(3):516–529.
- Debayan Banerjee, Sushil Awale, Ricardo Usbeck, and Chris Biemann. 2023. Dbp-quad: A question answering dataset over the dbp scholarly knowledge graph. *arXiv preprint arXiv:2303.13351*.
- Satanjeev Banerjee and Alon Lavie. 2005. **METEOR: An automatic metric for MT evaluation with improved correlation with human judgments**. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. **SciBERT: A pretrained language model for scientific text**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.
- Brad Botkin. 2024. **Usa basketball wins gold medal: Stephen curry, lebron james hold off france at 2024 paris olympics**. Accessed: 15.01.2025.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. 2024. **Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery**. *Preprint*, arXiv:2410.05080.

792	Cheng-Han Chiang and Hung-Yi Lee. 2023a. Are synonym substitution attacks really synonym substitution attacks? In <i>Findings of the Association for Computational Linguistics: ACL 2023</i> , pages 1853–1878.	848
793		849
794		850
795		851
796		852
797	Cheng-Han Chiang and Hung-yi Lee. 2023b. <a href="#">Can large language models be an alternative to human evaluations?</a> In <i>Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.	853
798		854
799		855
800		856
801		857
802		858
803	Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In <i>Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)</i> , pages 2924–2936.	859
804		860
805		861
806		862
807		
808		863
809		864
810		865
811	Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. <i>arXiv preprint arXiv:1803.05457</i> .	866
812		867
813		868
814		869
815		870
816	Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. <i>Advances in neural information processing systems</i> , 36:10088–10115.	871
817		872
818		873
819		874
820	Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. <a href="#">BERT: Pre-training of deep bidirectional transformers for language understanding</a> . In <i>Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)</i> , pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.	875
821		
822		876
823		877
824		878
825		879
826		880
827		
828		881
829	George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In <i>Proceedings of the Second International Conference on Human Language Technology Research, HLT '02</i> , page 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.	882
830		883
831		884
832		885
833		886
834		
835	Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2023. AlpacaFarm: A simulation framework for methods that learn from human feedback. <i>Advances in Neural Information Processing Systems</i> , 36:30039–30069.	887
836		888
837		889
838		890
839		891
840		892
841	Julia Evans, Jennifer D’Souza, and Sören Auer. 2024. Large language models as evaluators for scientific synthesis. In <i>Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)</i> , pages 1–22.	893
842		894
843		895
844		896
845		897
846	Jawwad Ahmed Farid. 2024. <a href="#">Can machines think?</a> Accessed: 15.01.2025.	898
847		899
	Jinlan Fu, See Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. Gptscore: Evaluate as you desire. In <i>Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)</i> , pages 6556–6576.	900
		901
	Siddhant Garg and Goutham Ramakrishnan. 2020. Bae: Bert-based adversarial examples for text classification. In <i>Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)</i> , pages 6174–6181.	902
		903
	Hamed Babaei Giglou, Jennifer D’Souza, and Sören Auer. 2024. Llm4synthesis: Leveraging large language models for scientific synthesis. <i>arXiv preprint arXiv:2409.18812</i> .	
	Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The false promise of imitating proprietary llms. <i>arXiv preprint arXiv:2305.15717</i> .	
	Jens Hauser, Zhao Meng, Damián Pascual, and Roger Wattenhofer. 2021. Bert is robust! a case against synonym-based adversarial examples in text classification. <i>arXiv preprint arXiv:2109.07403</i> .	
	Joey Hejna, Rafael Rafailov, Harshit Sikchi, Chelsea Finn, Scott Niekum, W Bradley Knox, and Dorsa Sadigh. 2023. Contrastive preference learning: Learning from human feedback without rl. <i>arXiv preprint arXiv:2310.13639</i> .	
	Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. <a href="#">Measuring massive multitask language understanding</a> . In <i>International Conference on Learning Representations</i> .	
	Robin Jia and Percy Liang. 2017. <a href="#">Adversarial examples for evaluating reading comprehension systems</a> . In <i>Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing</i> , pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.	
	Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In <i>Proceedings of the AAAI conference on artificial intelligence</i> , volume 34, pages 8018–8025.	
	Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In <i>Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)</i> , pages 2567–2577.	
	Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023. Prometheus: Inducing fine-grained evaluation	

904	capability in language models. In <i>The Twelfth International Conference on Learning Representations</i> .	
905		
906	Seungone Kim, Juyoung Suk, Shayne Longpre,	
907	Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham	
908	Neubig, Moontae Lee, Kyungjae Lee, and Minjoon	
909	Seo. 2024. Prometheus 2: An open source language	
910	model specialized in evaluating other language mod-	
911	els. <i>arXiv preprint arXiv:2405.01535</i> .	
912	Tom Kocmi and Christian Federmann. 2023. Large	
913	language models are state-of-the-art evaluators of	
914	translation quality. In <i>24th Annual Conference of the</i>	
915	<i>European Association for Machine Translation</i> , page	
916	193.	
917	Andreas Köpf, Yannic Kilcher, Dimitri von Rütte,	
918	Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens,	
919	Abdullah Barhoum, Duc Nguyen, Oliver Stan-	
920	ley, Richárd Nagyfi, et al. 2024. Openassistant	
921	conversations-democratizing large language model	
922	alignment. <i>Advances in Neural Information Process-</i>	
923	<i>ing Systems</i> , 36.	
924	Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit	
925	Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo.	
926	2023. Longeval: Guidelines for human evaluation of	
927	faithfulness in long-form summarization. In <i>Proceed-</i>	
928	<i>ings of the 17th Conference of the European Chap-</i>	
929	<i>ter of the Association for Computational Linguistics</i> ,	
930	pages 1650–1669.	
931	Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021.	
932	Hurdles to progress in long-form question answering.	
933	In <i>Proceedings of the 2021 Conference of the North</i>	
934	<i>American Chapter of the Association for Computa-</i>	
935	<i>tional Linguistics: Human Language Technologies</i> .	
936	Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Wein-	
937	berger. 2015. From word embeddings to document	
938	distances. In <i>International conference on machine</i>	
939	<i>learning</i> , pages 957–966. PMLR.	
940	VI Levenshtein. 1966. Binary codes capable of correct-	
941	ing deletions, insertions, and reversals. <i>Proceedings</i>	
942	<i>of the Soviet physics doklady</i> .	
943	Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bo-	
944	han Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang,	
945	and Huan Liu. 2025. Preference leakage: A contam-	
946	ination problem in llm-as-a-judge. <i>arXiv preprint</i>	
947	<i>arXiv:2502.01534</i> .	
948	Percy Liang, Rishi Bommasani, Tony Lee, Dimitris	
949	Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian	
950	Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Ku-	
951	mar, Benjamin Newman, Binhang Yuan, Bobby Yan,	
952	Ce Zhang, Christian Alexander Cosgrove, Christo-	
953	pher D Manning, Christopher Re, Diana Acosta-	
954	Navas, Drew Arad Hudson, Eric Zelikman, Esin	
955	Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren,	
956	Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel	
957	Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun,	
958	Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar	
959	Khattab, Peter Henderson, Qian Huang, Ryan An-	
960	drew Chi, Sang Michael Xie, Shibani Santurkar,	
	Surya Ganguli, Tatsunori Hashimoto, Thomas Icard,	961
	Tianyi Zhang, Vishrav Chaudhary, William Wang,	962
	Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Ko-	963
	reeda. 2023. <i>Holistic evaluation of language models</i> .	964
	<i>Transactions on Machine Learning Research</i> . Fea-	965
	tured Certification, Expert Certification.	966
	Chin-Yew Lin. 2004a. Rouge: A package for auto-	967
	matic evaluation of summaries. In <i>Text summariza-</i>	968
	<i>tion branches out</i> , pages 74–81.	969
	Chin-Yew Lin. 2004b. <i>ROUGE: A package for auto-</i>	970
	<i>matic evaluation of summaries</i> . In <i>Text Summariza-</i>	971
	<i>tion Branches Out</i> , pages 74–81, Barcelona, Spain.	972
	Association for Computational Linguistics.	973
	Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang,	974
	Yile Wang, and Yue Zhang. 2021. Logiqa: a	975
	challenge dataset for machine reading comprehen-	976
	sion with logical reasoning. In <i>Proceedings of the</i>	977
	<i>Twenty-Ninth International Conference on Interna-</i>	978
	<i>tional Joint Conferences on Artificial Intelligence</i> ,	979
	pages 3622–3628.	980
	Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang,	981
	Ruochen Xu, and Chenguang Zhu. 2023. G-eval:	982
	Nlg evaluation using gpt-4 with better human align-	983
	ment. In <i>Proceedings of the 2023 Conference on</i>	984
	<i>Empirical Methods in Natural Language Processing</i> ,	985
	pages 2511–2522.	986
	MetaAI. 2024. Introducing llama 3.1: Our most capa-	987
	ble models to date. <a href="https://ai.meta.com/blog/meta-llama-3-1/">https://ai.meta.com/blog/</a>	988
	<a href="https://ai.meta.com/blog/meta-llama-3-1/">meta-llama-3-1/</a> . Accessed: 16 February 2025.	989
	Ning Miao, Yee Whye Teh, and Tom Rainforth. 2024.	990
	<i>Selfcheck: Using LLMs to zero-shot check their own</i>	991
	<i>step-by-step reasoning</i> . In <i>The Twelfth International</i>	992
	<i>Conference on Learning Representations</i> .	993
	Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish	994
	Sabharwal. 2018. Can a suit of armor conduct elec-	995
	tricity? a new dataset for open book question an-	996
	swering. In <i>Proceedings of the 2018 Conference on</i>	997
	<i>Empirical Methods in Natural Language Processing</i> ,	998
	pages 2381–2391.	999
	MistralAI. 2024. Mistral large 2407. <a href="https://mistral.ai/en/news/mistral-large-2407">https://</a>	1000
	<a href="https://mistral.ai/en/news/mistral-large-2407">mistral.ai/en/news/mistral-large-2407</a> . Ac-	1001
	cessed: 16 February 2025.	1002
	Anastasios Nentidis, Georgios Katsimpras, Anasta-	1003
	sia Krithara, Salvador Lima-López, Eulàlia Farré-	1004
	Maduell, Martin Krallinger, Natalia Loukachevitch,	1005
	Vera Davydova, Elena Tutubalina, and Georgios	1006
	Paliouras. 2024a. Overview of bioasq 2024: the	1007
	twelfth bioasq challenge on large-scale biomedical	1008
	semantic indexing and question answering. In <i>Inter-</i>	1009
	<i>national Conference of the Cross-Language Evalua-</i>	1010
	<i>tion Forum for European Languages</i> , pages 3–27.	1011
	Springer.	1012
	Anastasios Nentidis, Georgios Katsimpras, Anastasia	1013
	Krithara, and Georgios Paliouras. 2024b. Overview	1014
	of bioasq tasks 12b and synergy12 in clef2024. <i>Work-</i>	1015
	<i>ing Notes of CLEF</i> , 2024.	1016



1017	Allard Oelen, Mohamad Yaser Jaradeh, Kheir Eddine Farfar, Markus Stocker, and Sören Auer. 2019. <a href="#">Comparing Research Contributions in a Scholarly Knowledge Graph</a> . In <i>Proceedings of the Third International Workshop on Capturing Scientific Knowledge co-located with the 10th International Conference on Knowledge Capture (K-CAP 2019)</i> , pages 21–26, Marina Del Rey, CA, USA.	1073
1018		1074
1019		1075
1020		
1021		1076
1022		1077
1023		1078
1024		1079
1025	Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. <i>Advances in neural information processing systems</i> , 35:27730–27744.	1080
1026		1081
1027		
1028		1082
1029		1083
1030		1084
1031	Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002a. Bleu: a method for automatic evaluation of machine translation. In <i>Proceedings of the 40th annual meeting of the Association for Computational Linguistics</i> , pages 311–318.	1085
1032		
1033		1086
1034		1087
1035		1088
1036	Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002b. <a href="#">Bleu: a method for automatic evaluation of machine translation</a> . In <i>Proceedings of the 40th Annual Meeting on Association for Computational Linguistics</i> , ACL '02, page 311–318, USA. Association for Computational Linguistics.	1089
1037		
1038		1090
1039		1091
1040		1092
1041		1093
1042	David Pride, Matteo Cancellieri, and Petr Knuth. 2023. Core-gpt: Combining open access research and large language models for credible, trustworthy question answering. In <i>International Conference on Theory and Practice of Digital Libraries</i> , pages 146–159. Springer.	1094
1043		
1044		1095
1045		1096
1046		1097
1047		1098
1048	Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: your language model is secretly a reward model. In <i>Proceedings of the 37th International Conference on Neural Information Processing Systems</i> , NIPS '23, Red Hook, NY, USA. Curran Associates Inc.	1099
1049		1100
1050		
1051		1101
1052		1102
1053		1103
1054		1104
1055	Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. <a href="#">SQuAD: 100,000+ questions for machine comprehension of text</a> . In <i>Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing</i> , pages 2383–2392, Austin, Texas. Association for Computational Linguistics.	1105
1056		
1057		1106
1058		1107
1059		1108
1060		1109
1061	Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. <i>Transactions of the Association for Computational Linguistics</i> , 7:249–266.	1110
1062		1111
1063		
1064		1112
1065	David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. <i>arXiv preprint arXiv:2311.12022</i> .	1113
1066		1114
1067		1115
1068		1116
1069		1117
1070	Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In <i>Proceedings of the 57th annual meeting of the association for computational linguistics</i> , pages 1085–1097.	1118
1071		1119
1072		1120
		1121
		1122
	Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In <i>Proceedings of the 2013 conference on empirical methods in natural language processing</i> , pages 193–203.	1123
		1124
		1125
		1126
		1127
		1128
		1129
	Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for nlg systems. <i>ACM Computing Surveys (CSUR)</i> , 55(2):1–39.	
	David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. 2019. <a href="#">Analysing mathematical reasoning abilities of neural models</a> . In <i>International Conference on Learning Representations</i> .	
	Amanpreet Singh, Mike D’Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2022. <a href="#">SciRepeval: A multi-format benchmark for scientific document representations</a> . In <i>Conference on Empirical Methods in Natural Language Processing</i> .	
	Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine de Marneffe, and Wei Xu. 2016. <a href="#">Results of the WNUT16 named entity recognition shared task</a> . In <i>Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)</i> , pages 138–144, Osaka, Japan. The COLING 2016 Organizing Committee.	
	Keh-Yih Su, Ming-Wen Wu, and Jing-Shin Chang. 1992. <a href="#">A new quantitative quality measure for machine translation systems</a> . In <i>COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics</i> .	
	Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In <i>Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)</i> , pages 641–651.	
	Minyang Tian, Luyu Gao, Dylan Zhang, Xinan Chen, Cunwei Fan, Xuefei Guo, Roland Haas, Pan Ji, Kittithat Krongchon, Yao Li, et al. 2024. Scicode: A research coding benchmark curated by scientists. In <i>The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track</i> .	
	Adam R Villafior, Zhe Huang, Swapnil Pande, John M Dolan, and Jeff Schneider. 2022. Addressing optimism bias in sequence modeling for reinforcement learning. In <i>international conference on machine learning</i> , pages 22270–22283. PMLR.	
	David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. <a href="#">Fact or fiction: Verifying scientific claims</a> . In <i>Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)</i> , pages 7534–7550, Online. Association for Computational Linguistics.	

1130	Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui	1188
1131	Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu,	1189
1132	and Jie Zhou. 2023a. Is chatgpt a good nlg evaluator?	1190
1133	a preliminary study. In <i>Proceedings of the 4th New</i>	1191
1134	<i>Frontiers in Summarization Workshop</i> , pages 1–11.	1192
1135	Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa	1193
1136	Liu, Noah A Smith, Daniel Khashabi, and Hannaneh	
1137	Hajishirzi. 2023b. Self-instruct: Aligning language	1194
1138	models with self-generated instructions. In <i>Proceed-</i>	1195
1139	<i>ings of the 61st Annual Meeting of the Association for</i>	1196
1140	<i>Computational Linguistics (Volume 1: Long Papers)</i> ,	1197
1141	pages 13484–13508.	1198
1142	Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni,	
1143	Abhranil Chandra, Shiguang Guo, Weiming Ren,	
1144	Aaran Arulraj, Xuan He, Ziyang Jiang, et al. 2024.	
1145	Mmlu-pro: A more robust and challenging multi-task	
1146	language understanding benchmark. <i>arXiv preprint</i>	
1147	<i>arXiv:2406.01574</i> .	
1148	Zhen Wang. 2022. Modern question answering	
1149	datasets and benchmarks: A survey. <i>arXiv preprint</i>	
1150	<i>arXiv:2206.15030</i> .	
1151	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten	
1152	Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou,	
1153	et al. 2022. Chain-of-thought prompting elicits reason-	
1154	ing in large language models. <i>Advances in neural</i>	
1155	<i>information processing systems</i> , 35:24824–24837.	
1156	Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan,	
1157	Lingfeng Shen, Benjamin Van Durme, Kenton Mur-	
1158	ray, and Young Jin Kim. 2024. Contrastive prefer-	
1159	ence optimization: pushing the boundaries of llm per-	
1160	formance in machine translation. In <i>Proceedings of</i>	
1161	<i>the 41st International Conference on Machine Learn-</i>	
1162	<i>ing</i> , ICML’24. JMLR.org.	
1163	Xi Yan, Patrick Westphal, Jan Seliger, and Ricardo Us-	
1164	beck. 2024. Bridging the gap: Generating a compre-	
1165	hensive biomedical knowledge graph question an-	
1166	swering dataset. In <i>ECAI 2024</i> , pages 1198–1205.	
1167	IOS Press.	
1168	Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeon-	
1169	bin Hwang, Seungone Kim, Yongrae Jo, James	
1170	Thorne, Juho Kim, and Minjoon Seo. 2024. <b>FLASK:</b>	
1171	<b>Fine-grained language model evaluation based on</b>	
1172	<b>alignment skill sets</b> . In <i>ICLR 2024 Workshop on</i>	
1173	<i>Large Language Model (LLM) Agents</i> .	
1174	Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q.	
1175	Weinberger, and Yoav Artzi. 2020. <b>Bertscore:</b>	
1176	<b>Evaluating text generation with bert</b> . <i>Preprint</i> ,	
1177	<i>arXiv:1904.09675</i> .	
1178	Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Chris-	
1179	tian M. Meyer, and Steffen Eger. 2019. <b>MoverScore:</b>	
1180	<b>Text generation evaluating with contextualized em-</b>	
1181	<b>beddings and earth mover distance</b> . In <i>Proceedings</i>	
1182	<i>of the 2019 Conference on Empirical Methods in</i>	
1183	<i>Natural Language Processing and the 9th Interna-</i>	
1184	<i>tional Joint Conference on Natural Language Pro-</i>	
1185	<i>cessing (EMNLP-IJCNLP)</i> , pages 563–578, Hong	
1186	Kong, China. Association for Computational Lin-	
1187	guistics.	
	Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan	1188
	Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin,	1189
	Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023.	1190
	Judging llm-as-a-judge with mt-bench and chatbot	1191
	arena. <i>Advances in Neural Information Processing</i>	1192
	<i>Systems</i> , 36:46595–46623.	1193
	Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B	1194
	Brown, Alec Radford, Dario Amodei, Paul Chris-	1195
	tiano, and Geoffrey Irving. 2019. Fine-tuning lan-	1196
	guage models from human preferences. <i>arXiv</i>	1197
	<i>preprint arXiv:1909.08593</i> .	1198
	<b>A Extended Related Work</b>	1199
	<b>A.1 Question &amp; Answering Benchmarks</b>	1200
	Automatic Q&A remains a rapidly evolving field,	1201
	with an expanding array of datasets supporting its	1202
	development. These datasets exhibit wide variation	1203
	in domains, question types, and generation method-	1204
	ologies. To contextualize the scienceQ&A focus of	1205
	this paper, we first provide an overview of related	1206
	datasets and highlight their distinctions.	1207
	A comprehensive review of NLP Q&A datasets	1208
	prior to 2022 by Wang (2022) outlines multi-	1209
	ple dataset categories. However, datasets using	1210
	multiple-choice answer formats fall outside the	1211
	scope of this work, including MCTest (2013) for fic-	1212
	tional stories, ARC (2018) for high-school science	1213
	exams, OpenBookQA (2018) for science facts, Pub-	1214
	MedQA (2019) for medical summarization, and	1215
	LogiQA (2021) for logical reasoning in exams.	1216
	Similarly, datasets with Boolean answer types (e.g.,	1217
	BoolQ (2019)) and numeric result datasets (e.g.,	1218
	Mathematics Dataset (2019)) are not within our	1219
	scope. Even datasets requiring entity-based an-	1220
	swers from structured knowledge sources, such as	1221
	ComplexWebQuestions (2018), diverge from our	1222
	focus on generative scienceQ&A. Among the 41 re-	1223
	viewed datasets, only the 2023 release of BioASQ	1224
	was found to be directly relevant to our research	1225
	objective.	1226
	Additional scientific Q&A datasets fall outside	1227
	our research scope, such as DBLP-QuAD (Baner-	1228
	jee et al., 2023) for bibliographic queries and	1229
	datasets designed to extract factual knowledge from	1230
	scientific literature for knowledge graph population	1231
	(Auer et al., 2023; Yan et al., 2024). One notable	1232
	dataset is DBLP-QuAD (Banerjee et al., 2023),	1233
	which contains 10,000 QA pairs generated via	1234
	SPARQL queries over the DBLP scholarly knowl-	1235
	edge graph (KG). While valuable for bibliographic	1236
	metadata QA, its focus is restricted to bibliographic	1237
	queries, limiting its applicability to more diverse	1238
	or conceptual scholarly questions. Another dataset	1239

we evaluated is the SciQA benchmark (Auer et al., 2023), which includes 100 handcrafted complex QA pairs alongside 2,465 automatically generated ones. These questions are derived from the Open Research Knowledge Graph (ORKG). However, the dataset’s reliance on ORKG-specific entities and the need for direct KG access to produce high-quality answers posed practical challenges for our study, leading us to exclude it. We also considered the BioKGQA dataset, proposed by Yan et al. (2024), which features 85,368 QA pairs generated using multi-noded triples from PrimeKG, a comprehensive KG oriented toward precision medicine. While its method of leveraging KG structures for QA generation is innovative, the dataset primarily focuses on fact-based answers, making it less suitable for exploring nuanced or multi-dimensional evaluation.

The Hugging Face leaderboard serves as a widely recognized benchmark space for new LLMs. Among its datasets, MMLU-PRO (Wang et al., 2024) provides expert-reviewed multiple-choice questions across diverse domains, including Medicine, Law, Engineering, and Mathematics. GPQA (Rein et al., 2023) similarly includes multiple-choice questions authored by domain experts in Biology, Physics, and Chemistry. However, these benchmarks primarily evaluate intrinsic reasoning ability in answering fixed-choice questions rather than generative Q&A tasks. Recent LLM evaluation trends emphasize human-aligned benchmarks for chat assistant alignment, such as MT-Bench and Chatbot Arena (Zheng et al., 2023), which assess open-domain, multi-turn dialogue abilities. These benchmarks evaluate capabilities in writing, reasoning, extraction, and domain knowledge (e.g., STEM and humanities/social sciences), but their focus remains distinct from our objective. Instead, our work aims to develop a framework that enables open-source LLMs to robustly evaluate generative scienceQ&A using standardized, rubrics-based multifaceted assessments, further pushing the frontier of LLM-as-a-judge capabilities.

## A.2 LLM-as-a-judge Evaluation Rubrics

The early notion of LLM-as-a-judge (Zheng et al., 2023) measured the correlation between an LLM judge and human evaluators in open-domain instruction following for NLG. Early works on using LLMs as evaluators of LLM-generated text emphasized pairwise evaluation along a single dimension

of ‘preference’ in determining which response was superior (Wang et al., 2023a; Chiang and Lee, 2023b; Dubois et al., 2023; Zheng et al., 2023; Liu et al., 2023; Kocmi and Federmann, 2023).

In these preference comparisons, evaluation rubrics emerged. G-Eval (Liu et al., 2023) incorporated criteria such as coherence, consistency, fluency, and relevance for summarization benchmarking. The rubrics of GPTScore (Fu et al., 2024) for summarization tasks also align with ours. However, GPTScore also includes dialogue rubrics aligned with instruction-following chat-based evaluations such as likeability, flexibility, inquisitiveness, and engagement, which are out of scope for scienceQ&A. Open-ended story generation rubrics (Chiang and Lee, 2023b) share similarities with scientific QA, where likeability is juxtaposed against informativeness, relevance to answer pertinence, and grammar to cohesion, integration, and readability.

ScienceAgentBench (Chen et al., 2024) employs four rubric-based metrics to evaluate LLMs and agentic AI for generating programmatic workflows suited to scientific data science applications. Their evaluation rubrics include Success Rate (SR) (task completion), Valid Execution Rate (VER) (program runs without error), CodeBERTScore (CBS) (similarity to reference implementations), and Cost Efficiency (measuring API expenses). They compiled data science benchmarks across Bioinformatics, Computational Chemistry, Geographical Information Science, and Psychology & Cognitive Neuroscience. Their work recommends automating rubric-based evaluations for generated code quality, a gap we address differently in this paper.

SciCode (Tian et al., 2024) sourced research-level coding problems across 16 subfields in natural science disciplines, benchmarking LLMs on pythonic code generation from docstring-format task descriptions. It relied on pure quantitative metrics such as pass@1 rate but lacked detailed rubric-based insights.

More recent works examine fine-grained rubrics. The FLASK (Ye et al., 2024) evaluation rubrics—Robustness, correctness, efficiency, factuality, commonsense, comprehension, insightfulness, completeness, metacognition, readability, conciseness, and harmlessness—are closely related to our work. Nine of their 12 rubrics align with ours, except insightfulness, harmlessness, and metacognition, which lack clear definitions for LLMs. Prometheus (Kim et al., 2023, 2024) devel-



Rubric	G-Eval (Liu et al., 2023)	GPTScore (Fu et al., 2024)	LLM-Alt (Chiang and Lee, 2023b)	FLASK (Ye et al., 2024)	Current Work
Coherence	✓	✓	✓		✓
Cohesion			✓		✓
Completeness		✓		✓	✓
Conciseness		✓		✓	✓
Correctness	✓ (Consistency)	✓ (Factuality)		✓ (Logical Correctness)	✓
Informativeness		✓		✓	✓
Integration					✓
Readability	✓ (Fluency)	✓	✓	✓	✓
Relevancy	✓	✓	✓		✓
Harmlessness				✓	
Logical Thinking				✓	
Insightfulness				✓	
Engagement		✓	✓		
Likeability		✓	✓		

Table 3: Comparison of evaluation rubrics across different works against ours (last column). Cells marked in green indicate rubrics that apply to a particular work.

opened 1,222 customized rubrics across open-domain benchmarks but relied on human reference answers. Instead, this work circumvents human annotation dependence by leveraging adversarial data where the desired behavior is known.

LLM-based scientific QA evaluation evolved from three criteria (Pride et al., 2023; Evans et al., 2024)—comprehensiveness, trust, and utility—to nine (Giglou et al., 2024), including relevancy, correctness, completeness, informativeness, integration, cohesion, readability, and conciseness.

HELM (Liang et al., 2023) set a precedent in holistic language model evaluation by considering seven quantitative metrics: accuracy, calibration, robustness, fairness, bias, toxicity, and efficiency. This multi-metric approach exposed trade-offs across different evaluation criteria, ensuring that accuracy was not the sole measure of performance. Inspired by this, our work targets a holistic evaluation of scienceQ&A, defining a multifaceted qualitative framework of nine rubrics that comprehensively assess LLM outputs.

This work advances the field by proposing a structured evaluation methodology that moves beyond traditional ROUGE/BLEU metrics and proprietary model dependencies. By equipping open-source LLMs with robust evaluation capabilities, we eliminate reliance on expensive human annotations and proprietary LLM-generated reference answers. Our approach ensures that the evaluation of scientific QA models remains transparent, replicable, and fine-grained, aligning with broader efforts to standardize LLM assessments.

## B Science Q&A Datasets

### B.1 The Multidisciplinary ORKGSyn Corpus

The ORKGSyn corpus is a highly multidisciplinary scienceQ&A dataset. In this context, Figure 4 illus-

trates the distribution of instances across 34 different scientific disciplines represented in ORKGSyn.

### B.2 The BioASQ Corpus

According to the scienceQ&A task definition, ORKGSyn consistently linked each  $Q$  to  $N = 5$  papers with abstracts, whereas in BioASQ,  $N$  varied between 1 and 40. Figure 5 presents the distribution of instances, categorized by the number of papers associated with each  $Q$ .

### B.3 Our Adversarial Corpus

By constructing an adversarial dataset, we introduce deliberate errors into the original outputs to test whether LLMs can detect and evaluate poor-quality responses. This allows for a comparative analysis of evaluation scores between original and manipulated datasets. While original syntheses may not always be flawless, adversarial datasets are designed to exhibit a marked deterioration in quality, and we expect evaluation scores to reflect this decline. We construct two tiers of adversarial datasets.

- Subtle Adversarial Dataset:** Here, reference texts are minimally altered, making it challenging for models to detect changes. These alterations mimic realistic errors that may go unnoticed in automated evaluations.
- Extreme Adversarial Dataset:** This dataset involves substantial modifications to reference texts, making the adversarial setting apparent and straightforward for models to identify. The evaluations should result in significantly lower scores.

Nine evaluation criteria are systematically targeted during adversarial dataset creation, with distinct



Figure 4: Number of Questions per Research Field on the ORKGSyn Dataset. The y-axis represents the "Research Fields".

manipulations tailored to degrade the corresponding aspect of synthesis quality. To simulate varying degrees of distortion, adversarial sentences were drawn from diverse sources: a blog post snippet from (Farid, 2024), a sentence from a sports news article provided by (Botkin, 2024), and a tweet sourced from (Strauss et al., 2016). Below, we detail the adversarial interventions and describe what each criterion evaluates:

1. **Relevancy:** Is the information in the answer relevant to the problem?

- Subtle: Append a sentence from a different synthesis paragraph within the same domain.
- Extreme: Append a sentence from an unrelated sports news article.

2. **Correctness:** Is the information in the answer a correct representation of the content of the provided abstracts?

- Subtle: Append a sentence from a different synthesis paragraph within the same domain.
- Extreme: Append a sentence from an unrelated sports news article.

3. **Completeness:** Is the answer a comprehensive encapsulation of the relevant information in the provided abstracts?

- Subtle: Remove the last sentence from the synthesis.
- Extreme: Remove the last sentence and append a sentence from an unrelated sports news article.

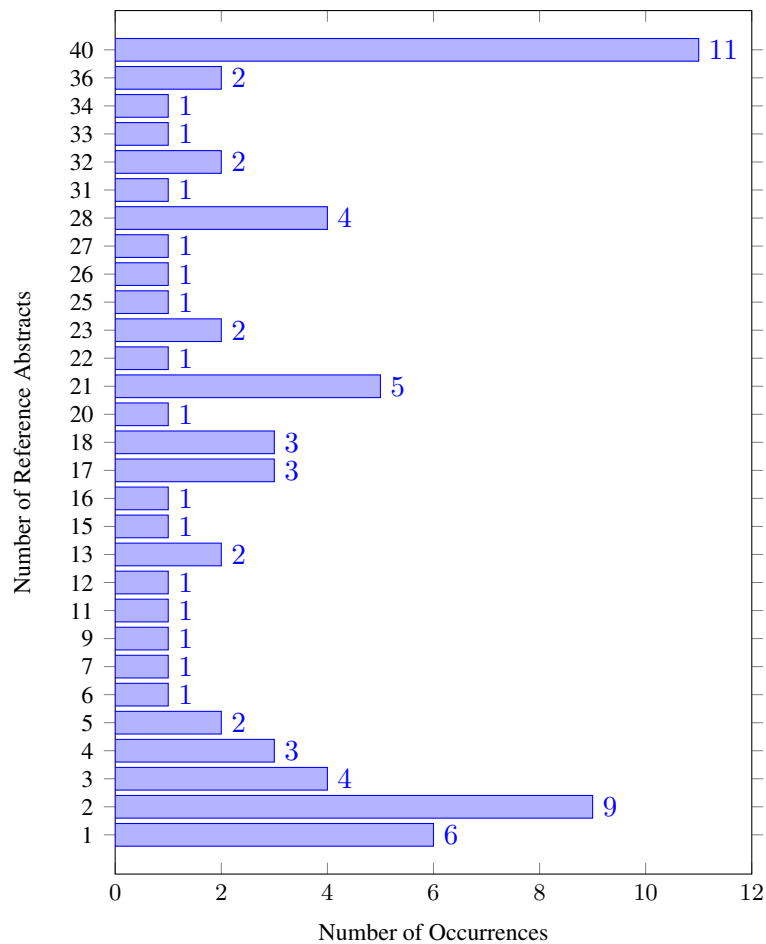


Figure 5: Number of Reference Abstracts per Question on the BioASQ dataset

4. **Informativeness:** Is the answer a useful and informative reply to the problem?

- Subtle: Append a sentence from a different synthesis paragraph within the same domain.
- Extreme: Append a sentence from an unrelated sports news article.

5. **Integration:** Are the sources structurally and linguistically well-integrated, using appropriate markers of provenance/quotation and logical connectors for each reference?

- Subtle: Remove the first logical connector (e.g., "however", "therefore").
- Extreme: Remove all logical connectors.

6. **Cohesion:** Are the sentences connected appropriately to make the resulting synthesis cohesive?

- Subtle: Swap the positions of the last two sentences.

– Extreme: Randomly shuffle all sentences.

7. **Coherence:** Are the ideas connected soundly and logically?

- Subtle: Append a sentence from a different synthesis paragraph within the same domain.
- Extreme: Append a sentence from an unrelated sports news article.

8. **Readability:** Does the answer follow appropriate style and structure conventions for academic writing, particularly for readability?

- Subtle: Append a snippet from a casual blog post.
- Extreme: Append a sentence from an informal tweet.

9. **Conciseness:** Is the answer short and clear, without redundant statements?



1483  
1484  
1485  
1486  
1487  
1488  
  
1489  
1490  
1491  
1492  
1493  
1494  
1495  
1496  
1497  
1498  
1499  
1500  
1501  
1502  
1503  
1504  
1505  
1506  
1507  
1508  
1509  
1510  
1511  
1512  
1513  
1514  
1515  
1516  
1517  
1518  
1519  
1520  
1521  
1522  
1523  
1524  
1525  
1526  
1527  
1528  
1529  
1530  
1531  
1532  
1533

- Subtle: Use the LLM to generate a redundant version of the last sentence and append it to the response.
- Extreme: Append a redundant version after every sentence in the original response.

For the conciseness criterion, redundant sentences were generated by prompting the original model to unnecessarily rephrase and extend its outputs. Despite specific instructions to avoid introductory phrases or comments about the redundancy task, the Llama models frequently failed to adhere to these guidelines, requiring extensive manual cleanup. For instance, phrases such as "Here is a redundant sentence..." were frequently appended to the generated responses. We manually revised over 100 responses generated by the Llama 8B model and 18 by the Llama 70B. This process culminated in the creation of adversarial datasets with subtle and extreme degrees of manipulation for each underlying dataset, resulting in a total of four datasets. By systematically introducing errors tailored to each quality criterion, these datasets enable a detailed evaluation of LLM performance across a range of distortions. This two-level adversarial approach provides a comprehensive method for assessing the models' sensitivity to various levels of quality deterioration and their ability to assign appropriate evaluation scores under varying conditions.

One of the early papers on LLM-as-a-judge (Zheng et al., 2023) also used the strategy of an adversarial attack to test whether the judge LLM could detect verbosity bias which is related to our attack for redundancy.

A key NLP line of research in text adversarial attacks involves synonym substitution attacks (SSAs) (Alzantot et al., 2018) generate adversarial samples by replacing words with synonyms in benign text, relying on sophisticated methods such as TextFooler (Jin et al., 2020), PWWS (Ren et al., 2019), and BAE (Garg and Ramakrishnan, 2020). While SSAs were conjectured to produce low-quality text in fluency and meaning (Hauser et al., 2021; Chiang and Lee, 2023a), there was no guarantee of this, necessitating human raters to compare their assessments against LLM scores. In contrast, our adversarial attacks, ranging from subtle to extreme, deliberately violate English syntax, ensuring inherently low-quality text and obviating the need for human raters. This design directly

tests whether the LLM can overcome its optimism bias and robustly assign the expected low scores. 1534  
1535

**C Manual Subsample Observations of Vanilla LLM<sub>eval</sub>** 1536  
1537

We conducted a manual analysis of evaluation scores generated by Llama 8B, using a structured annotation process to assess the model's performance. The human adjudicator categorized issues based on six criteria from (Kim et al., 2023): "rejected feedback is not consistent with its score," "too general and abstract," "overly optimistic," "not relevant to the response," "overly critical," and "unrelated to the score rubric." This analysis covered 20 questions from ORKGsyn and 10 from BioASQ. 1538  
1539  
1540  
1541  
1542  
1543  
1544  
1545  
1546  
1547  
1548

Our findings reveal that Llama 8B's evaluations of vanilla syntheses are generally aligned with human judgement. However, the feedback provided by the model tends to be overly general and abstract. For instance, while Llama 8B often identifies a lack of minor details, it frequently fails to specify what exactly is missing. In the adversarial settings, Llama 8B exhibits overly optimistic scoring across all syntheses. This pattern is consistent for both subtle and extreme adversarial datasets, though the ORKGsyn dataset receives even higher scores compared to BioASQ. Additionally, we observed that in a small fraction of cases, the scores were not relevant to the response. For example, Llama 8B occasionally hallucinated its own evaluation criteria and scored based on these fabricated metrics. Another notable issue is the inconsistency between the rationale provided and the assigned scores. In a few cases, the evaluation feedback explicitly states that there is almost no unnecessary information in the synthesis, yet the model assigns a perfect score of 5. Furthermore, for a significant portion of the scores, the rationale was very general and abstract, often merely reciting the evaluation characteristic guidelines provided in the system prompt without offering specific insights. 1549  
1550  
1551  
1552  
1553  
1554  
1555  
1556  
1557  
1558  
1559  
1560  
1561  
1562  
1563  
1564  
1565  
1566  
1567  
1568  
1569  
1570  
1571  
1572  
1573  
1574

**D Detailed Quantitative Experimental Results** 1575  
1576

To obtain a comprehensive assessment of the LLM's performance on the scientific QA task, we conducted a quantitative analysis of their outputs. This evaluation leverages *summarisation-based* (such as BLUE (Papineni et al., 2002b), ROUGE (Lin, 2004b), METEOR (Banerjee and 1577  
1578  
1579  
1580  
1581  
1582

Lavie, 2005), NIST (Doddington, 2002), and BERTScore (Zhang et al., 2020)) and *edit distance-based* (such as WER (Su et al., 1992), WMD (Kusner et al., 2015), and MoverScore (Zhao et al., 2019)) metrics. These metrics measure the similarity between text outputs, providing a quantitative estimate of how closely the generated responses align with one another. The evaluation process involves treating each LLM response as a reference and pairing it with the outputs of every other model as candidates, resulting in all possible pairwise combinations. This approach ensures that we capture not only the absolute performance of each model but also their relative alignment. By examining the similarity scores between models, we gain insights into how consistently they generate responses.

### D.1 Summarisation-Based Metrics

Evaluating the quality of generated text often requires comparing it to reference summaries. To achieve this, summarisation-based evaluation metrics measure the degree of textual overlap between the generated and reference texts, typically assessing aspects such as precision, recall, and n-gram similarity. These metrics are widely used in NLP tasks such as machine translation, text summarisation, and question answering. The summarisation-based metrics (i.e. BLUE, ROUGE, METEOR, NIST, and BERTScore) are described as follows.

**Bilingual Evaluation Understudy (BLEU).** The BLEU (Papineni et al., 2002b) metric automates machine translation evaluation, offering a cost-effective alternative to human assessment. It measures n-gram overlap between a candidate and reference text, with a brevity penalty to prevent short translations. While widely used, BLEU has limitations, including insensitivity to semantic meaning and struggles with short texts. According to the Figure 6, an anomaly occurs for the *LLaMA-3.1-70B* model on the BioASQ dataset, where the score is 98.63. This discrepancy arises because the BLUE metric averages over n-gram scores, including cases where higher-order n-grams (e.g., four-grams) are absent in shorter outputs, forcing those scores to zero. Interestingly, the results reveal patterns of correlation among specific models. For instance, *Mistral-Large* and *Qwen2.5-72B* exhibit moderate alignment, with scores around 25, indicating some similarity in their outputs. On the ORKGSynthesis dataset, the *LLaMA-3.1-70B* and *LLaMA-3.1-8B* models display a relatively high

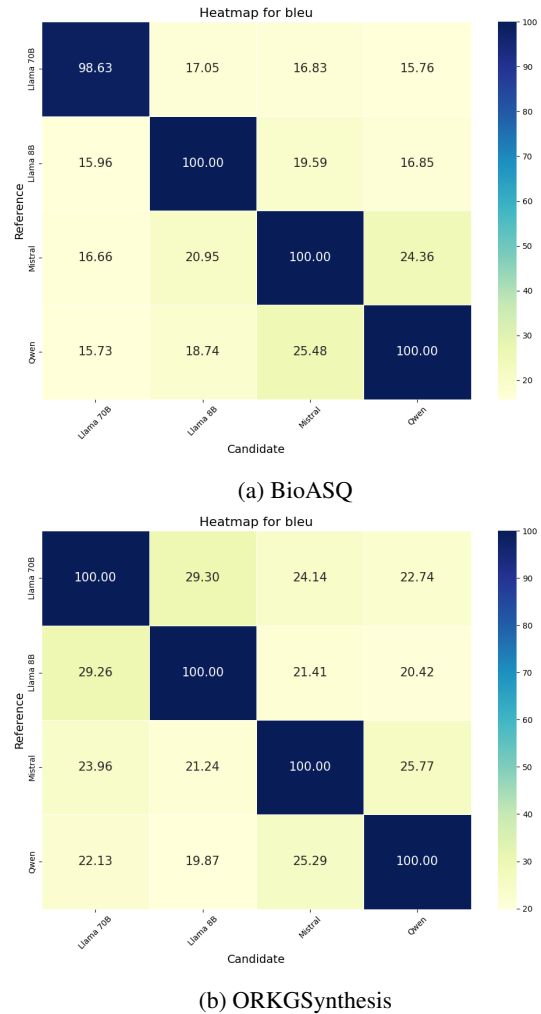
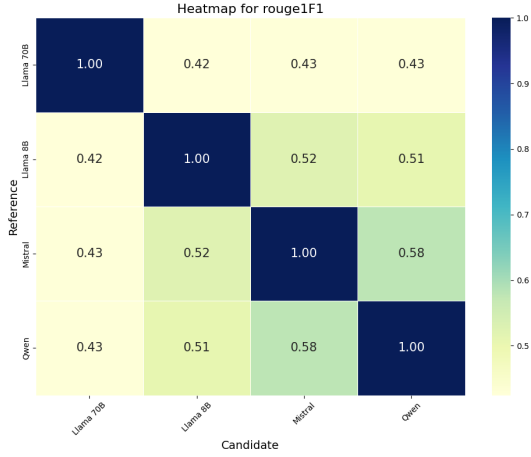


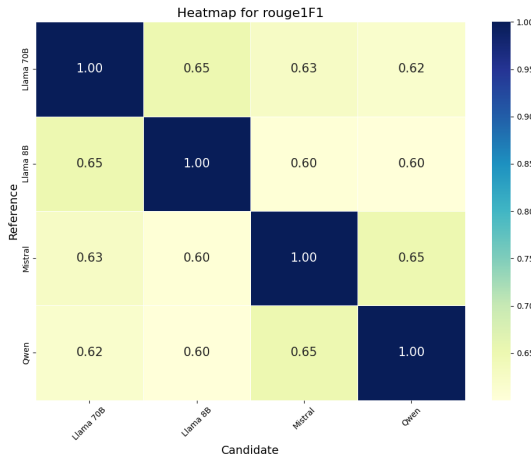
Figure 6: BLEU metric results.

correlation, likely due to their shared Meta origin, similar training data, and primary differences in parameter count. In terms of individual performance, *Mistral-Large* achieves the highest average BLUE score on the BioASQ dataset, while on the ORKGSynthesis dataset, Meta’s 70B LLM performs best.

**Recall-Oriented Understudy for Gisting Evaluation (ROUGE).** ROUGE (Lin, 2004b) is a recall-based metric contrasting with BLEU’s precision focus. It measures n-gram, word sequence, and word pair overlap between machine-generated and human summaries. ROUGE-N emphasizes n-gram recall, while ROUGE-L captures the longest common subsequence (LCS) for better semantic similarity. Variants like ROUGE-W reward consecutive matches, and ROUGE-S use skip-bigrams for F-measure calculation. For this analysis, we used ROUGE-1 for unigram overlap, reflecting term alignment and content coverage, and ROUGE-L

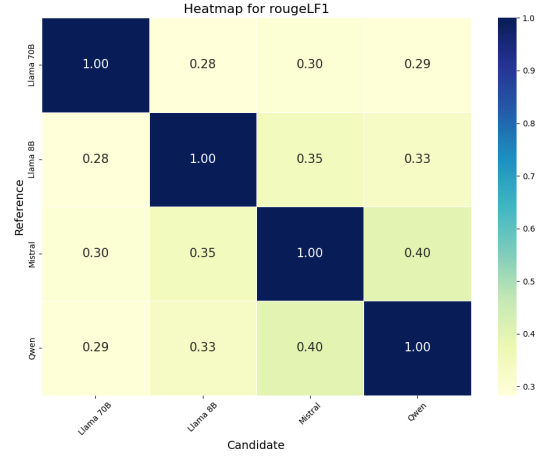


(a) BioASQ

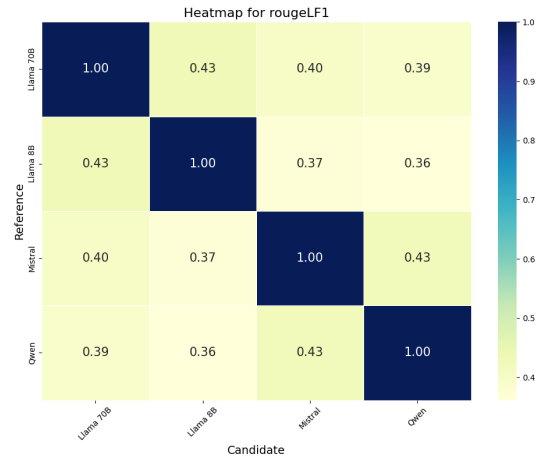


(b) ORKGSynthesis

Figure 7: ROUGE-1 metric results.



(a) BioASQ



(b) ORKGSynthesis

Figure 8: ROUGE-L metric results.

to assess sentence-level structure and coherence.

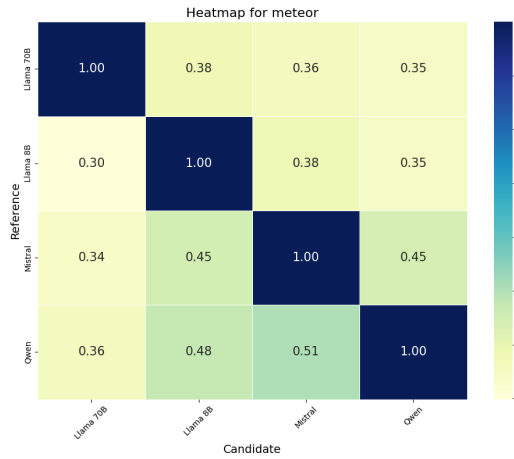
Illustrated results for ROUGE-1 in Figure 7 and ROUGE-L in Figure 8 exhibit distinct patterns across the two datasets. On the BioASQ dataset, *Mistral-Large* and *Qwen2.5-72B* demonstrate a notable correlation with a ROUGE-1 score of 0.58, suggesting shared thematic elements or vocabulary usage in their outputs. Additionally, a ROUGE-L score of 0.4 indicates a moderate similarity in sentence structure and coherence. In contrast, the *LLaMA-3.1-70B* model performs relatively poorly, which may stem from inconsistencies in generating high-quality outputs for certain tasks in this domain. On the ORKGSynthesis dataset, the *LLaMA-3.1-70B* model achieves the highest correlation, particularly when compared to the smaller LLaMA model. Mistral and *Qwen2.5-72B* also display a strong correlation on this dataset, reinforcing their observed alignment. Exclusively focusing on unigrams in ROUGE-1 yields a higher correlation than BLUE

and ROUGE-L, where overlaps are analyzed up to the four-gram level. This is because unigram-based evaluations inherently capture a broader overlap by disregarding strict positional constraints or dependencies on higher-order matches.

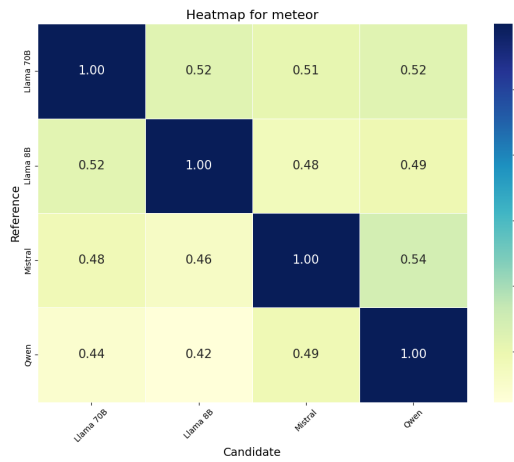
**Metric for Evaluation of Translation with Explicit Ordering (METEOR).** METEOR (Banerjee and Lavie, 2005) improves upon BLEU by prioritizing recall, which better aligns with human judgments. Unlike BLEU’s brevity penalty, METEOR explicitly integrates recall into its scoring. It also replaces BLEU’s reliance on higher-order n-grams with direct word alignment, enhancing semantic and structural accuracy. Additionally, METEOR avoids BLEU’s zero-score issue by using an alignment-based approach that captures partial matches and syntactic nuances more effectively.

Results showed in Figure 9 for METEOR metric, which produces scores ranging from 0 (no correlation) to 1 (perfect correlation). In our evaluation,



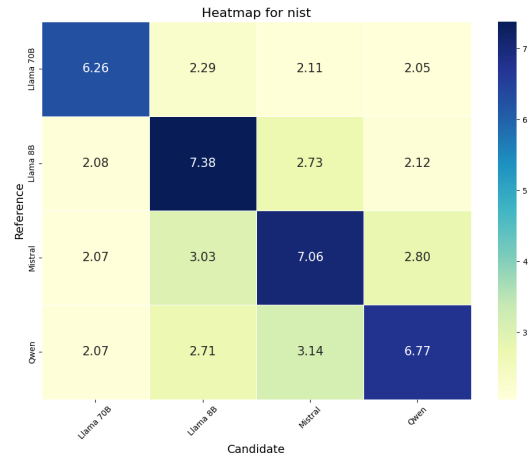


(a) BioASQ

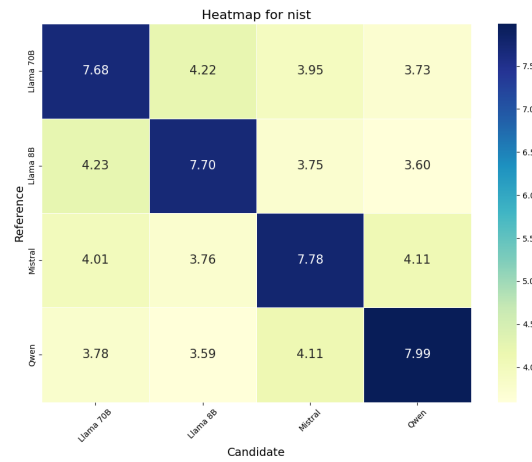


(b) ORKGSynthesis

Figure 9: METEOR metric results.



(a) BioASQ



(b) ORKGSynthesis

Figure 10: NIST metric results

METEOR highlighted differences between the models. On the BioASQ dataset, the *LLaMA-3.1-70B* model performed poorly, consistent with earlier metrics. Conversely, the *LLaMA-3.1-8B* model achieved comparable scores to *Mistral-Large* and *Qwen2.5-72B*, suggesting it generates responses that align well with unigram matches. For the ORKGSynthesis dataset, the scores were more uniform across models, reflecting a general similarity in performance. However, an intriguing pattern emerged with *Qwen2.5-72B*: when used as a reference, its scores varied substantially compared to when it was a candidate. This discrepancy may be attributable to METEOR’s emphasis on recall. Higher scores when *Qwen2.5-72B* is a candidate suggest it produces longer outputs, increasing the likelihood of matches with reference terms. This raises questions about the relationship between generation length and perceived quality in evaluation metrics, which warrants further investigation.

**US National Institute of Standards and Technology (NIST).** The NIST (Doddington, 2002) score builds on the BLUE metric but introduces a significant enhancement by focusing on the information content of n-grams. Unlike BLUE, which traditionally weights all matches equally, NIST prioritizes matches with greater informational value. This approach stems from the observation that rarer words or n-grams convey more specific and meaningful information than frequent ones. Consequently, matching infrequent n-grams contributes more to the NIST score, making it a more context-sensitive evaluation metric.

The NIST scores are unbounded and range from 0 (indicating poor quality) to higher positive values, with larger scores reflecting better-quality matches. Unlike BLUE, the NIST score’s sensitivity to informational content allows it to provide more nuanced assessments. As shown in Figure 10 for the BioASQ dataset, the *LLaMA-3.1-70B* model

consistently underperforms, even when compared against its output. This poor performance can be attributed to the model’s tendency to generate shorter responses and repeat words within a single response, leading to lower overall information gain. In contrast, other models demonstrate relatively strong correlations, suggesting more balanced and information-rich outputs. In the ORKGSynthesis dataset, correlations are notably higher across all models. The *LLaMA-3.1-70B* model achieves strong alignment with its smaller counterpart, consistent with observations in other metrics. Similarly, *Mistral-Large* and *Qwen2.5-72B* continue to display strong correlations.

**BERTScore.** BERTScore (Zhang et al., 2020) is designed to evaluate text similarity by leveraging contextualized embeddings from pre-trained transformer models like BERT (Devlin et al., 2019). Unlike traditional metrics that rely on exact token matches or n-gram overlap, BERTScore assesses semantic similarity at a token level, enabling it to recognize paraphrases, capture long-range dependencies, and account for nuanced semantic ordering. This approach allows it to surpass earlier metrics in evaluating complex and varied outputs, as it does not depend solely on surface-level text similarity. The strength of BERTScore lies in its ability to align tokens in a candidate sentence with those in a reference sentence using their embeddings, which encode rich contextual information. Studies by Zhang et al. (2020) demonstrate that BERTScore highly correlates with human judgment, making it a valuable tool for evaluating machine-generated text.

BERTScore generates values between 0 and 1, where a score closer to 1 indicates stronger semantic alignment. In our analysis w.r.t Figure 11, BERTScore highlights the limitations of the *LLaMA-3.1-70B* model on the BioASQ dataset, as its tendency to produce incomplete or incoherent answers results in lower scores. The other models, including those evaluated on the ORKGSynthesis dataset, achieve scores near 0.9, which reflects a high degree of semantic similarity between their outputs. Unlike metrics such as BLUE, ROUGE, or METEOR, which primarily evaluate surface-level similarity based on token overlap or n-gram matches, BERTScore incorporates the contextual meaning of tokens. This capability allows it to capture deeper semantic relationships, even when lexical or syntactic differences exist between candidate and reference sentences. Consequently, while ear-

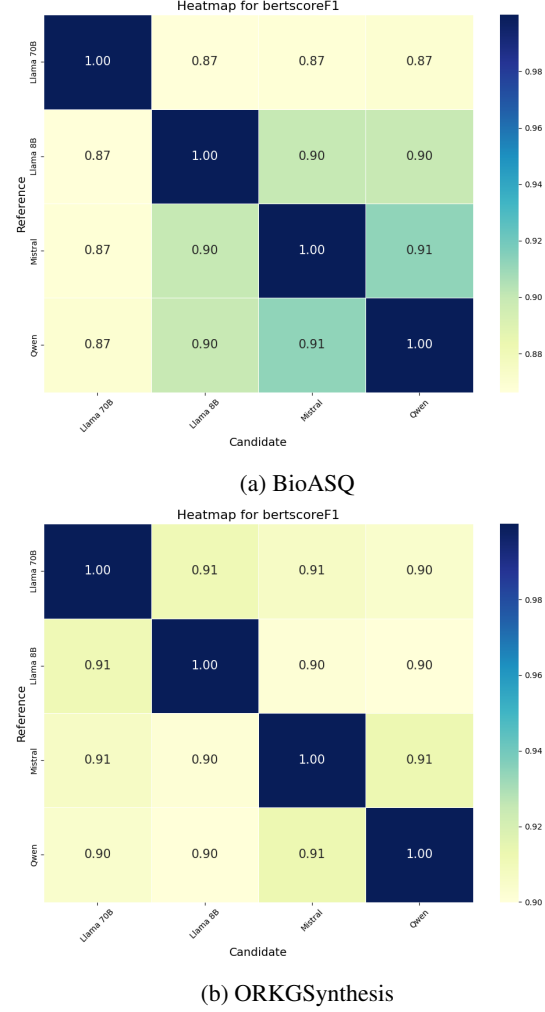


Figure 11: BERTScore metric results.

lier metrics may indicate lower correlation, particularly for models generating paraphrased or stylistically varied outputs, BERTScore reveals that the underlying semantic content remains closely aligned. This observation underscores the potential of embedding-based metrics to provide a more nuanced evaluation of language model outputs, particularly in tasks where paraphrasing and creative rewording are common. Furthermore, the high BERTScore results across most models suggest that their outputs are semantically coherent, even if traditional metrics fail to capture this aspect. This finding highlights the value of embedding-based metrics in complementing traditional approaches, providing a broad evaluation framework. Future work could explore fine-tuning the embeddings used in BERTScore to align even more closely with domain-specific human judgments, particularly in specialized tasks such as medical or scientific text generation.

## D.2 Edit Distance-Based Metrics

To complement summary-based evaluation, we also incorporated edit distance-based metrics, which assess text similarity by measuring the number of modifications required to convert one string into another. This approach originates from Levenshtein distance (Levenshtein, 1966), a fundamental concept in text processing commonly applied in spell-checking and auto-correction. Edit distance evaluation quantifies text dissimilarity through insertions, deletions, and substitutions. For instance, converting the word *mug* to *hut* requires substituting two characters (*m* to *h* and *g* to *t*), which, if we assign a cost of 1 per insertion/deletion and 2 per substitution, yields an edit distance of 4. These costs can be adjusted depending on the task.

**Word Error Rate (WER).** The WER (Su et al., 1992) is a similarity metric grounded in the concept of minimum edit distance, measuring the number of edits (substitutions, insertions, and deletions) required to transform a candidate text into reference text. Unlike character-level edit distance, WER operates at the word level, treating entire words as the basic transformation units. It is particularly suited for speech recognition and machine translation, where word-level alignment is essential. However, the metric has inherent limitations, particularly its sensitivity to word order. Sentences with semantically identical meanings but different word arrangements are heavily penalised, potentially leading to overly pessimistic assessments of similarity.

The analysis of WER according to Figure 12 scores across different datasets and models reveals several notable patterns. For the BioASQ dataset, the *LLaMA-3.1-70B* model exhibits markedly poor performance as a reference. This behavior can be attributed to the model’s tendency to produce shorter outputs. Since WER normalizes the number of edits by reference length, shorter reference texts amplify the impact of any discrepancies, leading to inflated WER values. This observation aligns with earlier findings highlighting *LLaMA-3.1-70B*’s challenges in generating comprehensive responses for the BioASQ dataset. In contrast, for the ORKGSynthesis dataset, *Qwen2.5-72B* demonstrates superior performance when serving as a reference. However, as a candidate, *Qwen2.5-72B* achieves lower scores. This discrepancy likely arises from *Qwen2.5-72B*’s tendency to generate longer sequences. In such cases, the normalization

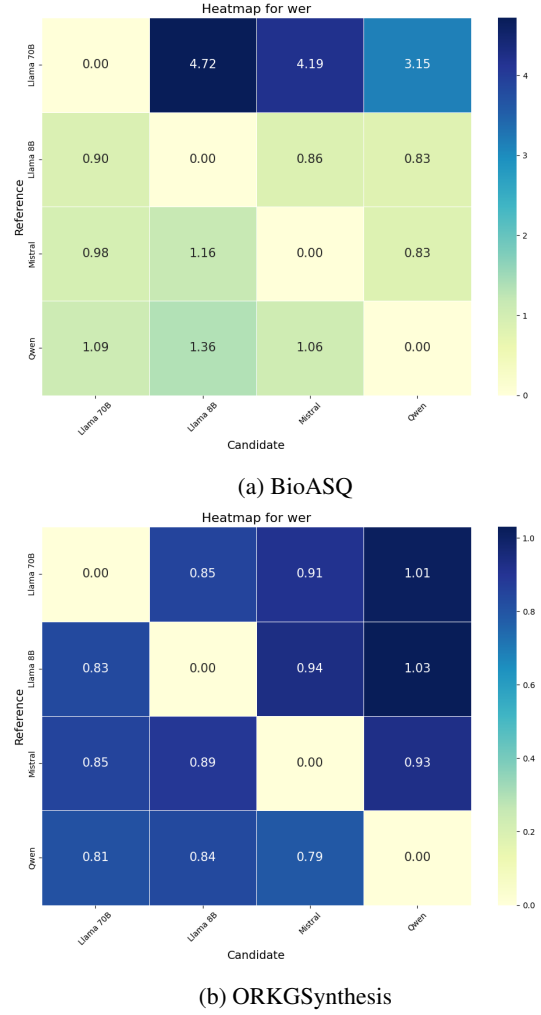


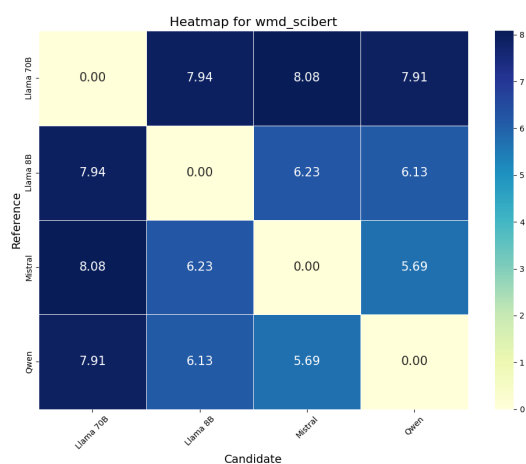
Figure 12: WER metric results

by reference length in the WER formula leads to more significant deviations when *Qwen2.5-72B*’s outputs are compared against shorter references. These findings underscore the metric’s dependence on the relative length of candidate and reference texts, which can introduce biases when evaluating models with different generation strategies. While WER provides a straightforward measure of surface-level similarity, its inability to account for semantic equivalence or tolerate variations in word order limits its applicability in evaluating generative models. For instance, outputs with paraphrased structures or stylistic differences might receive high WER scores despite being semantically aligned with the reference.

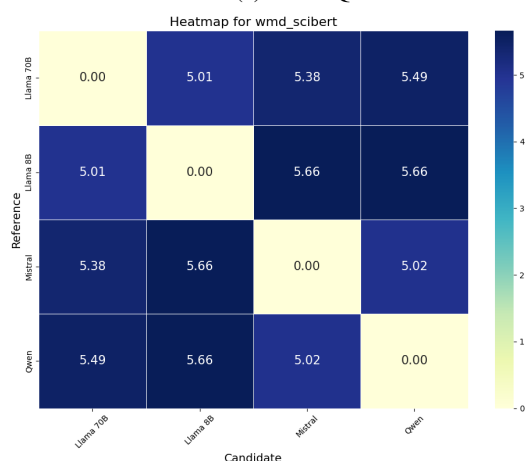
**Word Mover’s Distance (WMD).** The WMD (Kusner et al., 2015) quantifies the dissimilarity between two text documents by calculating the minimum cumulative "distance" that the embedded words in one document must travel to align



with the words in another document. The metric is inspired by the Earth Movers Distance, a concept in optimal transport theory, which measures the minimum work required to transform one probability distribution into another. Unlike greedy matching approaches like BERTScore, WMD leverages an optimal matching strategy, ensuring a more precise alignment of semantically relevant terms. In its original formulation, WMD used Word2Vec embeddings to represent words as vectors in a continuous space. However, we employed SPECTER2 (Singh et al., 2022) and SciBERT (Beltagy et al., 2019) embeddings, designed for scientific texts and have shown superior performance in domain-specific applications.



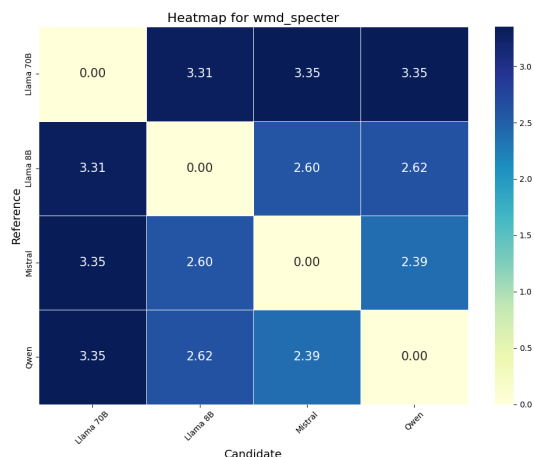
(a) BioASQ



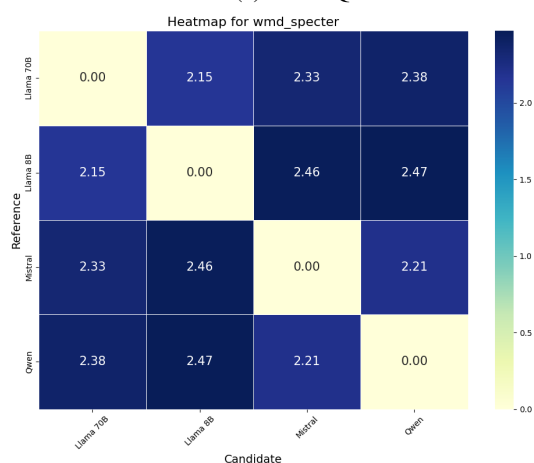
(b) ORKGSynthesis

Figure 13: WMD metric results using SciBERT as embeddings.

The results of WMD using SPECTER2 embeddings are represented in Figure 14 and WMD with SciBERT is represented in Figure 13. A WMD score of 0 indicates perfect alignment between the



(a) BioASQ



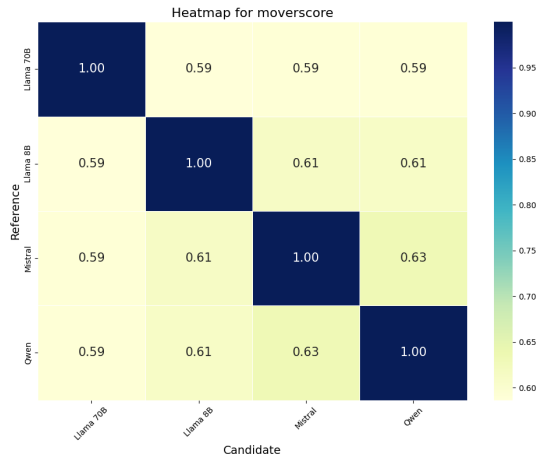
(b) ORKGSynthesis

Figure 14: WMD metric results using SPECTER2 as embeddings.

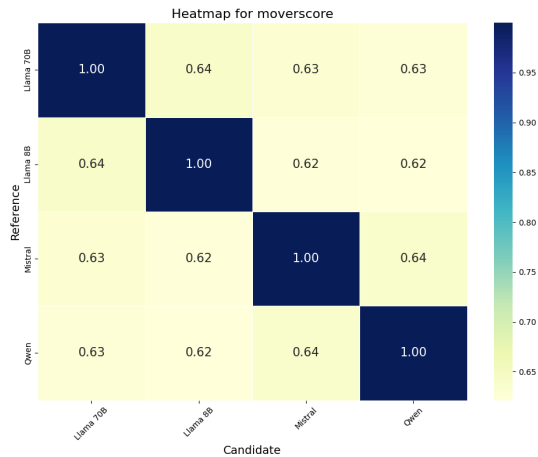
candidate and reference texts, while higher scores denote greater dissimilarity. Analysis of the WMD scores across datasets highlights several trends. As expected, *LLaMA-3.1-70B* exhibits poor performance on the BioASQ dataset, consistent with its tendency to generate incoherent responses. In contrast on the ORKGSynthesis dataset, the two LLaMA models show strong alignment with one another, as do *Mistral-Large* and *Qwen2.5-72B*. An intriguing finding is the significant difference in WMD scores when using SPECTER embeddings compared to SciBERT embeddings. Specifically, SPECTER embeddings yield substantially lower WMD scores, suggesting they provide better semantic representations for this task.

**MoverScore.** The MoverScore (Zhao et al., 2019) is an advanced metric that extends the principles of WMD to evaluate the dissimilarity of text documents by comparing both words and n-grams. One

of its primary advantages lies in its use of contextual embeddings, such as those generated by BERT, instead of static embeddings. This enables MoverScore to capture nuanced meanings, including word sense disambiguation and contextual relationships. Another key improvement is its allowance for many-to-one soft alignments, enabling more flexible matching between text elements. Furthermore, MoverScore incorporates inverse document frequency (IDF) weighting, emphasising rare and meaningful words, ensuring that these words contribute more significantly to the similarity score. The combination of BERT’s contextual embeddings and IDF weighting has been shown by Zhao et al. (2019) to correlate highly with human judgment.



(a) BioASQ



(b) ORKGSynthesis

Figure 15: MoverScore metric results

MoverScore produces a value between 0 and 1, where higher scores correspond to greater similarity between the candidate and reference texts. Across both datasets as shown in Figure 15, Mover-

Score results hover around 0.62, reflecting moderate alignment. As observed with other metrics, the *LLaMA-3.1-70B* model performs slightly worse on the BioASQ dataset, likely attributable to the model’s struggles with generating comprehensive and coherent responses. In contrast, the ORKGSynthesis dataset reveals strong correlations between the outputs of the *Mistral-Large* and *Qwen2.5-72B* models, and between the LLaMA models. The relatively consistent scores across models suggest that, according to MoverScore, the quality of responses is comparable. This uniformity can be interpreted as evidence that the models generate outputs with similar semantic content and structure, despite potential stylistic differences.

## E Evaluation Scoring Rubrics and the Five-point Likert Scale

This appendix presents the quality rubrics and their corresponding 5-point Likert scale descriptions used in our evaluation. Table 4 outlines the Linguistic and Stylistic quality rubrics, assessing aspects such as cohesion, conciseness, and readability. Table 5 details the Logical and Structural Integrity rubrics, evaluating factors like argument coherence, integration, and relevancy. Lastly, Table 6 describes the Content Accuracy and Informativeness rubrics, which measure correctness, completeness, and informativeness.

## F Experimental Setup

**Vanilla  $LLM_{eval}$  models.** The Vanilla  $LLM_{eval}$  models employ various LLMs, including *Mistral-Large-Instruct*, *LLaMA-3.1-70B-Instruct*, *Qwen2.5-72B-Instruct*, and *LLaMA-3.1-8B-Instruct*, as  $LLM_{gen}$  to generate response to the questions  $Q$  based on provided relevant papers. The evaluations for these models are conducted using  $LLM_{eval}$ , where each LLM acts as a generator and evaluator in a pairwise format. For example, *Mistral-Large-Instruct* serves as the evaluator for all four  $LLM_{gen}$  models, and the same process is applied to the other LLMs. The evaluation results are rated on a 5-point Likert scale, with each rating accompanied by a rationale that explains the model’s reasoning. This configuration serves as a baseline for comparing the performance of different  $LLM_{eval}$ .

**SFT (benign).** The *SFT (benign)* experiment involves finetuning the *LLaMA-3.1-8B-Instruct* model using structured response  $A$ ’s from all four

Rubric	5-point Likert scale description
<b>Cohesion</b>	<p><i>Rating 1. Very bad:</i> The sentences within the synthesis are disconnected, resulting in a disjointed and fragmented narrative.</p> <p><i>Rating 2. Bad:</i> There are attempts at connecting sentences, but the synthesis often feels disjointed.</p> <p><i>Rating 3. Moderate:</i> The sentences are connected in a way that the synthesis is mostly cohesive, with some areas of improvement.</p> <p><i>Rating 4. Good:</i> The synthesis is cohesive, with sentences well-connected to form a unified narrative.</p> <p><i>Rating 5. Very good:</i> The synthesis is highly cohesive, with all sentences and paragraphs logically connected, facilitating a clear and coherent narrative flow.</p>
<b>Conciseness</b>	<p><i>Rating 1. Very Bad:</i> The synthesis is verbose and cluttered with redundant or irrelevant information, significantly detracting from its clarity and focus.</p> <p><i>Rating 2. Bad:</i> The synthesis includes some redundant or irrelevant statements, detracting from its clarity.</p> <p><i>Rating 3. Moderate:</i> The synthesis is relatively clear and to the point, but could be more concise by eliminating a few redundant elements.</p> <p><i>Rating 4. Good:</i> The synthesis is concise and to the point, with virtually no redundant statements or unnecessary information.</p> <p><i>Rating 5. Very Good:</i> The synthesis is precisely concise, delivering information clearly and directly without any superfluous details or redundancy, enhancing its clarity and impact.</p>
<b>Readability</b>	<p><i>Rating 1. Very bad:</i> The synthesis is poorly written, with pervasive issues in style, structure, and language use, making it difficult to understand.</p> <p><i>Rating 2. Bad:</i> The text has noticeable issues with style, structure, or language use, affecting clarity.</p> <p><i>Rating 3. Moderate:</i> The synthesis follows appropriate conventions and uses language correctly, with minor issues in style or structure.</p> <p><i>Rating 4. Good:</i> The text is well-structured and easy to read, with language that is appropriately used and only minor stylistic improvements needed.</p> <p><i>Rating 5. Very good:</i> The synthesis is exceptionally well-written, following stylistic and structural conventions with precise language use, making it accessible and easy to read.</p>

Table 4: **Linguistic and Stylistic** Quality Rubrics and their 5-point Likert Scale Description

Rubric	5-point Likert scale description
<b>Coherence</b>	<p><i>Rating 1. Very bad:</i> The synthesis lacks logical connection between ideas, leading to a narrative that is confusing and difficult to follow.</p> <p><i>Rating 2. Bad:</i> The ideas are not always logically connected, leading to a somewhat confusing narrative.</p> <p><i>Rating 3. Moderate:</i> The ideas are logically connected for the most part, but the narrative could be strengthened for better clarity.</p> <p><i>Rating 4. Good:</i> The ideas are logically and soundly connected, offering a clear and understandable narrative.</p> <p><i>Rating 5. Very good:</i> The ideas within the synthesis are connected in a logical and sound manner, forming a coherent and compelling narrative that is easy to follow.</p>
<b>Integration</b>	<p><i>Rating 1. Very Bad:</i> The synthesis fails to integrate the sources in any meaningful way. It lacks appropriate markers, connectors, or transitions between ideas and fails to combine the information into a single, cohesive paragraph.</p> <p><i>Rating 2. Bad:</i> The sources are somewhat integrated but inconsistently. The use of markers and connectors is sporadic or inappropriately applied, and the information is presented in multiple paragraphs without a clear unifying structure.</p> <p><i>Rating 3. Moderate:</i> The sources are integrated into a coherent manner within one or multiple paragraphs. The transitions or connections could be smoother, and the text would benefit from better paragraph structure to enhance clarity and unity.</p> <p><i>Rating 4. Good:</i> The sources are well-integrated, using appropriate markers and connectors to create a seamless narrative. The information is effectively organized into a single paragraph, showing a clear, unified approach.</p> <p><i>Rating 5. Very Good:</i> The synthesis seamlessly integrates information from the various sources, using appropriate markers and connectors to create a smooth and unified narrative. All information is skillfully condensed into a single, well-structured paragraph, exemplifying excellent integration.</p>
<b>Relevancy</b>	<p><i>Rating 1. Very bad:</i> The information provided does not relate to the research question, showing a lack of understanding or connection to the topic.</p> <p><i>Rating 2. Bad:</i> The information occasionally relates to the research question but lacks direct and consistent relevance.</p> <p><i>Rating 3. Moderate:</i> The information is generally related to the research question, with occasional lapses in direct relevance.</p> <p><i>Rating 4. Good:</i> The information is consistently relevant to the research question, with only minor exceptions.</p> <p><i>Rating 5. Very good:</i> The synthesis is directly and consistently relevant to the research question, demonstrating a deep understanding of the topic and its nuances.</p>

Table 5: **Logical and Structural Integrity** Quality Rubrics and their 5-point Likert Scale Description



Rubric	5-point Likert scale description
<b>Correctness</b>	<i>Rating 1. Very bad:</i> The synthesis consistently misrepresents or inaccurately portrays the content of the provided abstracts, showing a significant deviation from the original sources. <i>Rating 2. Bad:</i> The synthesis contains several inaccuracies or misinterpretations of the source abstracts. <i>Rating 3. Moderate:</i> The synthesis accurately represents most of the content from the provided abstracts but may contain minor errors. <i>Rating 4. Good:</i> The synthesis provides an accurate representation of the content from the provided abstracts with minor exceptions. <i>Rating 5. Very good:</i> The information in the synthesis is an accurate and faithful representation of the content from the provided abstracts, without any factual errors or misinterpretations.
<b>Completeness</b>	<i>Rating 1. Very bad:</i> The synthesis omits most of the relevant information, failing to capture the essential points or details from the provided abstracts. <i>Rating 2. Bad:</i> Significant portions of relevant information from the provided abstracts are missing. <i>Rating 3. Moderate:</i> The synthesis captures a fair amount of the relevant information, though it may overlook some details. <i>Rating 4. Good:</i> The synthesis includes almost all relevant information, missing only minor details. <i>Rating 5. Very good:</i> The synthesis comprehensively encapsulates all relevant information from the provided abstracts, leaving no pertinent details or points unaddressed.
<b>Informativeness</b>	<i>Rating 1. Very bad:</i> The synthesis offers no valuable insights or useful information in response to the research question, lacking depth and utility. <i>Rating 2. Bad:</i> The answer provides limited new insights or useful information in response to the research question. <i>Rating 3. Moderate:</i> The answer is somewhat informative, offering insights or useful information but not in a comprehensive or detailed manner. <i>Rating 4. Good:</i> The answer is informative and insightful, providing comprehensive information in response to the research question. <i>Rating 5. Very good:</i> The synthesis is highly informative, providing valuable insights and detailed information that thoroughly addresses the research question.

Table 6: **Content Accuracy and Informativeness** Quality Rubrics and their 5-point Likert Scale Description

models from  $x = LLM_{gen}$  as inputs and quality assessments from  $y = LLM_{eval}$  as outputs. The aim is to fine-tune  $LLM_{eval}$  and this process is conducted using benign datasets, which contain no adversarial examples, to ensure that the model is trained in a controlled, non-hostile environment. The goal is to refine the model’s performance in a straightforward, non-challenging scenario.

**SFT (benign) + RL (adversarial).** In this experiment, the finetuned *LLaMA-3.1-8B-Instruct* model is further finetuned using the RLAIIF technique, focusing solely on adversarial datasets. As seen in the results from Figure 3, while *SFT (benign)* learns how to be the evaluator, but fails in recognizing the bad examples in adversarial sets. The *SFT (benign) + RL (adversarial)* helps  $LLM_{eval}$  to imitate the bad examples as well.

**SFT (benign) + RL (benign + adversarial).** Considering, the *SFT (benign) + RL (adversarial)* experiments, while the model achieves high performance in identifying adversarial samples, it struggles to assign higher Likert scale ratings to benign samples. This is because the model’s threshold for defining a "good" sample has been lowered to a rating of 3. The reason for such a drop in the maximum Likert scale in testing is due to poor penalization in RL that considered the desirable threshold of  $\leq 3$  during the sampling of the data

#### Conciseness Evaluation

```
{"Readability": { "rating": 1, "rationale": "The synthesis is poorly written, with pervasive issues in style, structure, and language use, making it difficult to understand due to the inclusion of an unrelated joke about #jets fans and #marksanchez at the end." } }
```

Figure 16: The  $y$  format for fine-tuning models.

preparation for RLAIIF for the subtle adversarial set. Essentially, training the model exclusively on adversarial behavior leads to confusion when it encounters benign examples. To solve this, the *SFT (benign) + RL (benign + adversarial)* is ideal for imitating both benign and adversarial behaviors, which ideally leads to success.

## G Training Details

**Supervised Fine-Tuning.** The fine-tuning threats each  $LLM_{gen}$  as an input  $x$  and outputs of all four  $LLM_{eval}$  as a  $y$ . The fine-tuning *SFT* dataset stats are represented in Table 2. It includes research questions, synthesized responses, source papers, and evaluation rubrics scale to ensure the model

learns both synthesis assessment and evaluation reasoning. The Figure 16 shows the  $y = LLM_{eval}$  format during the fine-tuning process for the "Readability" rubric (the example is chosen from the Adversarial set). The tokenized sequences were processed with a maximum sequence length specific to each dataset (ORKG-Synthesis: 4,645 tokens, BioASQ: 8,874 tokens). We finetuned the model per dataset for 5 epochs using the Paged AdamW 8-bit optimizer, ensuring memory efficiency. Key hyperparameters included a batch size of 1 per GPU (we used two GPUs), gradient accumulation steps equal to batch size, a learning rate of  $2e - 4$  with a warmup ratio of 0.03, weight decay of 0.001, and a max gradient norm of 0.3. The fine-tuned model checkpoints were saved for further analysis and RL-based fine-tuning.

**Reinforcement Learning.** Several hyperparameters and configurations for training an LLM with RL are used. Key hyperparameters include a learning rate of  $2e - 4$ , a batch size of 1 (per GPU), and a total of 2 training epochs. The model's training is configured with a maximum prompt length of 4500 and a maximum completion length of 150. The CPO uses a per-device batch size of 1, gradient accumulation steps of 1, and mixed precision (fp16). Additionally, the model undergoes fine-tuning with a learning rate of  $2e - 4$ .

**Hardware and Resource Allocation.** For the system setup, two H100 GPUs were utilized for RL fine-tunings, each with 80 GB of GPU memory, while for SFT models, only one H100 GPU with the same memory capacity was used. The CPU configuration for SFT involved 60 GB of memory with 8 cores of CPU, while RL took 60 GB of memory and 16 CPU cores.

## H Detailed Qualitative Experimental Results

This section represents the detailed qualitative experimental results for seven models. In the tables,  $LLM_{eval}$  models are defined as follows in the table columns:

- **M1:** Qwen2.5-72B
- **M2:** LLaMA-3.1-70B
- **M3:** Mistral-Large
- **M4:** LLaMA-3.1-8B
- **M5:** SFT (benign)

- **M6:** SFT (benign) + RL (adversarial)
- **M7:** SFT (benign) + RL (benign + adversarial)

The results for the BioASQ dataset are presented in Table 7 for LLaMA-3.1-8B  $LLM_{gen}$ , Table 8 for LLaMA-3.1-70B  $LLM_{gen}$ , Table 9 for Qwen2.5-72B  $LLM_{gen}$ , and Table 10 for Mistral-Large  $LLM_{gen}$ . While, for the ORKGSynthesis dataset, results are presented in Table 11 for LLaMA-3.1-8B  $LLM_{gen}$ , Table 12 for LLaMA-3.1-70B  $LLM_{gen}$ , Table 13 for Qwen2.5-72B  $LLM_{gen}$ , and Table 14 for Mistral-Large  $LLM_{gen}$ .

## I Example Evaluations

This section presents example evaluations of LLM-generated responses based on different rubrics. The examples were selected randomly from experimental models  $LLM_{eval}$  outputs in test sets. The examples showcase both well-performing and problematic cases, emphasizing the strengths and weaknesses of different models and fine-tuning strategies. The examples are shown in Table 15 and Table 16, which illustrate the challenges in vanilla and SFT models in distinguishing the desired behaviors. However, the *SFT (benign) + RL (benign+adversarial)* model demonstrates superior performance, even in terms of its rationales.

Examples from BioASQ in Table 15 show that *SFT (benign) + RL (adversarial)* often struggle with coherence and factual accuracy, as evidenced in examples 3 and 5. In contrast, models combining *SFT (benign) + RL (benign + adversarial)* tend to deliver more relevant and informative outputs, as seen in example 4. In Table 16 for the ORKGSynthesis dataset, examples in extreme settings, such as the *SFT (benign) + RL (adversarial)* rationales in example 8, demonstrate that  $LLM_{eval}$  specifically identifies the inclusion of off-topic sentences, resulting in lower relevance and completeness ratings, as anticipated.

Set	Rubrics	M1	M2	M3	M4	M5	M6	M7
benign	1. Coherence	4.95	4.91	4.73	5.00	4.95	3.00	4.91
	2. Cohesion	4.95	4.91	4.68	4.95	4.95	3.00	4.68
	3. Completeness	4.41	4.33	4.27	4.91	4.23	3.00	4.32
	4. Conciseness	4.95	4.91	4.68	4.45	4.77	3.00	3.82
	5. Correctness	4.95	4.92	4.64	4.73	5.00	3.00	5.00
	6. Informativeness	4.82	4.82	4.86	5.00	4.68	3.00	5.00
	7. Integration	4.91	4.91	4.55	4.77	4.95	3.00	4.59
	8. Readability	4.95	4.91	4.73	4.64	4.95	3.00	4.55
	9. Relevancy	4.95	4.83	4.86	5.00	5.00	3.00	5.00
extreme	1. Coherence	2.50	3.00	3.00	4.64	3.55	1.00	1.00
	2. Cohesion	4.05	4.55	3.82	4.77	4.93	3.00	4.41
	3. Completeness	2.68	3.31	3.27	4.64	3.82	1.05	1.00
	4. Conciseness	1.41	1.00	1.50	2.27	4.77	1.00	1.00
	5. Correctness	3.91	3.90	3.73	4.68	3.59	1.00	1.00
	6. Informativeness	3.59	4.20	3.68	4.68	3.41	1.09	1.00
	7. Integration	3.73	4.08	3.18	4.77	4.82	3.00	3.14
	8. Readability	2.05	2.15	2.55	4.14	1.89	1.00	1.00
	9. Relevancy	2.64	3.40	3.55	4.86	2.77	1.00	1.00
subtle	1. Coherence	3.82	4.10	4.14	4.82	4.86	2.91	2.73
	2. Cohesion	4.05	4.25	4.23	4.68	4.82	3.00	4.64
	3. Completeness	4.14	4.09	3.82	4.82	4.36	3.00	3.91
	4. Conciseness	3.09	2.60	2.41	3.68	4.55	2.41	2.27
	5. Correctness	4.23	4.50	4.14	4.64	4.86	3.00	2.82
	6. Informativeness	4.18	4.40	4.23	4.95	4.59	3.00	2.77
	7. Integration	4.09	4.27	3.86	4.64	4.91	3.00	4.64
	8. Readability	3.14	2.92	3.45	4.18	4.18	2.18	1.82
	9. Relevancy	4.32	4.20	4.36	5.00	4.77	2.77	2.73

Table 7: BioASQ dataset detailed evaluation results of various  $LLM_{eval}$  over  $LLM_{gen} = \text{LLaMA-3.1-8B}$ .

Set	Rubrics	M1	M2	M3	M4	M5	M6	M7
benign	1. Coherence	2.91	4.90	2.91	4.91	3.64	2.55	3.09
	2. Cohesion	2.91	4.90	2.86	4.86	4.05	2.32	2.86
	3. Completeness	2.50	3.29	2.73	4.64	3.86	2.45	2.64
	4. Conciseness	2.91	4.80	3.36	4.23	3.73	2.00	2.50
	5. Correctness	3.00	3.73	3.18	4.86	3.91	2.91	3.09
	6. Informativeness	2.77	4.36	3.00	4.95	3.27	2.50	2.95
	7. Integration	2.86	4.55	2.77	4.77	3.86	1.95	2.82
	8. Readability	2.91	4.90	3.05	4.41	3.23	2.05	2.86
	9. Relevancy	2.95	3.80	3.23	4.95	3.73	2.50	3.09
extreme	1. Coherence	1.50	2.09	2.14	3.59	2.73	1.00	1.00
	2. Cohesion	2.50	3.92	2.64	4.50	3.45	2.27	2.68
	3. Completeness	1.73	2.36	2.32	4.18	2.59	1.09	1.00
	4. Conciseness	1.50	1.17	1.55	2.41	3.91	1.18	1.00
	5. Correctness	2.32	3.27	2.50	3.91	2.95	1.00	1.00
	6. Informativeness	2.09	3.18	2.41	3.95	2.05	1.00	1.00
	7. Integration	2.23	3.38	2.32	4.45	3.14	1.91	2.14
	8. Readability	1.50	1.75	1.50	3.68	2.05	1.00	1.00
	9. Relevancy	1.45	2.73	2.41	3.95	2.00	1.00	1.00
subtle	1. Coherence	1.95	3.40	2.32	4.41	3.09	2.45	1.91
	2. Cohesion	2.50	3.33	2.45	4.64	3.55	2.32	2.95
	3. Completeness	2.59	3.36	2.95	4.77	3.59	2.64	2.27
	4. Conciseness	1.86	1.86	1.64	3.23	3.82	1.64	1.32
	5. Correctness	2.27	3.50	2.82	4.55	3.50	3.00	2.18
	6. Informativeness	2.32	3.82	2.59	4.55	3.14	2.68	2.18
	7. Integration	2.45	3.83	2.50	4.36	3.41	1.86	2.86
	8. Readability	1.91	1.92	2.41	4.05	3.27	1.68	1.50
	9. Relevancy	2.36	3.67	2.68	4.82	3.14	2.41	2.09

Table 8: BioASQ dataset detailed evaluation results of various  $LLM_{eval}$  over  $LLM_{gen} = \text{LLaMA-3.1-70B}$ .



Set	Rubrics	M1	M2	M3	M4	M5	M6	M7
benign	1. Coherence	5.00	5.00	4.95	4.95	5.00	3.00	4.95
	2. Cohesion	5.00	5.00	4.91	4.95	5.00	3.00	4.45
	3. Completeness	4.27	4.00	4.36	4.73	4.27	3.00	4.05
	4. Conciseness	5.00	4.91	4.91	4.55	4.86	3.00	4.45
	5. Correctness	4.95	5.00	4.91	4.82	4.91	3.00	4.91
	6. Informativeness	4.68	4.83	4.91	4.82	4.86	3.00	4.95
	7. Integration	4.91	5.00	4.86	4.77	4.91	3.00	4.86
	8. Readability	5.00	5.00	5.00	4.45	4.86	3.00	4.86
	9. Relevancy	5.00	5.00	5.00	4.91	5.00	3.00	4.91
extreme	1. Coherence	2.45	2.50	2.82	4.36	3.32	1.00	1.00
	2. Cohesion	4.09	4.46	3.77	4.73	4.86	3.00	4.59
	3. Completeness	2.32	3.10	2.68	3.95	3.91	1.00	1.00
	4. Conciseness	2.27	1.60	2.09	3.27	4.77	1.09	1.09
	5. Correctness	3.68	3.70	3.77	4.36	3.36	1.00	1.00
	6. Informativeness	3.41	3.60	3.50	4.41	2.32	1.00	1.00
	7. Integration	3.59	4.00	3.23	4.68	4.95	3.00	3.14
	8. Readability	1.91	1.75	2.23	4.23	2.05	1.00	1.00
	9. Relevancy	2.18	2.70	3.59	4.68	2.55	1.00	1.00
subtle	1. Coherence	3.64	4.10	4.09	4.77	4.91	3.00	3.00
	2. Cohesion	4.18	4.45	4.09	4.68	4.86	3.00	4.59
	3. Completeness	3.64	4.10	3.86	4.95	4.27	3.00	3.59
	4. Conciseness	2.82	2.70	2.36	3.86	4.59	2.59	2.14
	5. Correctness	4.23	4.30	4.00	4.59	4.73	3.00	3.14
	6. Informativeness	4.14	4.50	4.18	4.82	4.64	3.00	3.18
	7. Integration	4.00	4.55	4.00	4.86	4.95	3.00	4.45
	8. Readability	2.55	2.45	3.36	4.09	4.23	2.05	1.91
	9. Relevancy	4.23	4.50	4.27	4.86	4.95	2.95	3.00

Table 9: BioASQ dataset detailed evaluation results of various  $LLM_{eval}$  over  $LLM_{gen} = \text{Qwen2.5-72B}$ .

Set	Rubrics	M1	M2	M3	M4	M5	M6	M7
benign	1. Coherence	4.91	5.00	4.95	5.00	5.00	3.00	4.91
	2. Cohesion	4.91	4.92	4.77	5.00	4.95	3.00	4.73
	3. Completeness	4.27	4.00	4.41	4.57	4.32	3.00	4.36
	4. Conciseness	4.95	4.67	4.95	4.59	4.73	2.86	3.77
	5. Correctness	4.95	5.00	4.82	4.86	4.95	3.00	4.95
	6. Informativeness	4.64	4.75	4.91	4.91	4.73	3.00	4.91
	7. Integration	4.86	4.83	4.64	5.00	4.95	3.00	4.68
	8. Readability	4.95	4.83	5.00	4.50	4.75	2.91	4.86
	9. Relevancy	4.91	4.93	4.95	5.00	5.00	3.00	5.00
extreme	1. Coherence	2.27	2.09	3.14	4.41	3.50	1.00	1.00
	2. Cohesion	3.95	4.42	3.77	4.68	4.91	3.00	4.18
	3. Completeness	2.32	2.45	2.91	4.18	3.68	1.00	1.00
	4. Conciseness	1.95	1.38	1.82	3.09	4.45	1.09	1.09
	5. Correctness	3.68	3.45	3.73	4.55	3.41	1.00	1.00
	6. Informativeness	3.45	3.18	3.73	4.45	3.00	1.05	1.00
	7. Integration	3.59	3.77	3.09	4.68	4.82	2.91	3.00
	8. Readability	2.05	2.15	2.27	4.00	2.14	1.00	1.00
	9. Relevancy	2.32	2.92	3.64	4.73	3.32	1.00	1.00
subtle	1. Coherence	3.77	3.80	4.14	4.91	4.91	2.95	3.05
	2. Cohesion	4.00	4.79	4.27	4.77	4.86	3.00	4.41
	3. Completeness	3.86	4.20	3.50	4.91	4.50	3.00	3.68
	4. Conciseness	2.91	3.15	2.82	3.91	4.91	2.59	2.27
	5. Correctness	4.23	4.27	4.14	4.77	4.68	3.00	3.23
	6. Informativeness	4.18	4.20	4.32	4.86	4.59	2.95	3.14
	7. Integration	3.91	4.09	3.77	4.59	4.95	3.00	4.45
	8. Readability	2.82	2.54	3.73	4.36	4.59	2.27	1.95
	9. Relevancy	4.27	4.18	4.41	5.00	5.00	2.95	2.95

Table 10: BioASQ dataset detailed evaluation results of various  $LLM_{eval}$  over  $LLM_{gen} = \text{Mistral-Large}$ .

Set	Rubrics	M1	M2	M3	M4	M5	M6	M7
benign	1. Coherence	4.74	4.95	4.78	4.97	4.95	3.00	4.98
	2. Cohesion	4.72	4.93	4.71	4.97	4.88	3.00	4.95
	3. Completeness	4.28	4.40	4.52	4.50	4.59	3.00	4.76
	4. Conciseness	4.74	4.71	4.70	4.18	4.77	3.00	4.03
	5. Correctness	4.86	4.97	4.79	4.83	4.91	3.00	4.94
	6. Informativeness	4.57	4.93	4.77	4.97	4.89	3.06	4.97
	7. Integration	4.70	4.90	4.68	4.89	4.93	3.00	4.91
	8. Readability	4.77	4.94	4.86	4.41	4.85	3.00	4.86
	9. Relevancy	4.70	4.92	4.81	4.99	4.93	3.02	4.96
extreme	1. Coherence	2.58	3.42	3.35	4.27	4.81	1.04	1.00
	2. Cohesion	3.76	4.33	3.44	4.85	4.78	3.00	4.41
	3. Completeness	2.63	2.94	3.56	3.69	4.59	1.00	1.00
	4. Conciseness	1.11	1.05	1.57	1.70	4.64	1.00	1.00
	5. Correctness	3.81	4.09	3.98	4.24	4.82	1.01	1.00
	6. Informativeness	3.62	4.26	3.80	4.78	4.82	1.03	1.00
	7. Integration	3.37	3.90	3.16	4.72	4.86	2.73	3.10
	8. Readability	2.20	2.13	2.93	3.77	4.76	1.01	1.00
	9. Relevancy	2.78	3.51	3.58	4.99	4.81	1.02	1.00
subtle	1. Coherence	4.21	4.70	4.51	4.81	4.95	3.00	3.44
	2. Cohesion	3.98	4.49	3.96	4.90	4.84	3.00	4.52
	3. Completeness	4.10	4.23	3.92	4.99	4.67	3.00	4.20
	4. Conciseness	2.71	2.36	2.50	3.62	4.70	2.47	2.38
	5. Correctness	4.48	4.78	4.44	4.57	4.93	3.00	3.37
	6. Informativeness	4.35	4.79	4.50	4.90	4.87	3.03	3.38
	7. Integration	4.01	4.52	4.02	4.92	4.91	2.99	4.72
	8. Readability	3.22	2.88	3.79	4.49	4.82	2.85	2.98
	9. Relevancy	4.35	4.70	4.56	4.99	4.96	2.99	3.54

Table 11: ORKGSynthesis dataset detailed evaluation results of various  $LLM_{eval}$  over  $LLM_{gen} = \text{LLaMA-3.1-8B}$ .

Set	Rubrics	M1	M2	M3	M4	M5	M6	M7
benign	1. Coherence	4.77	4.97	4.78	5.00	4.88	3.00	4.98
	2. Cohesion	4.76	4.95	4.65	4.99	4.87	3.00	4.98
	3. Completeness	4.32	4.46	4.54	4.61	4.69	3.00	4.75
	4. Conciseness	4.77	4.75	4.65	4.15	4.73	2.99	4.05
	5. Correctness	4.89	4.96	4.83	4.80	4.92	3.00	4.92
	6. Informativeness	4.65	4.92	4.77	4.98	4.87	3.01	4.95
	7. Integration	4.76	4.90	4.56	4.92	4.92	3.00	4.96
	8. Readability	4.80	4.98	4.90	4.41	4.89	3.00	4.94
	9. Relevancy	4.73	4.90	4.75	4.99	4.93	3.03	4.94
extreme	1. Coherence	2.66	3.39	3.33	4.31	4.82	1.01	1.00
	2. Cohesion	3.80	4.30	3.56	4.88	4.87	3.00	4.46
	3. Completeness	2.74	2.97	3.56	3.63	4.56	1.00	1.00
	4. Conciseness	1.09	1.02	1.62	2.29	4.56	1.00	1.00
	5. Correctness	3.90	4.07	3.85	4.25	4.89	1.02	1.00
	6. Informativeness	3.76	4.30	3.87	4.69	4.81	1.02	1.00
	7. Integration	3.50	4.03	3.27	4.74	4.88	2.83	3.07
	8. Readability	2.16	2.03	2.95	3.85	4.71	1.00	1.00
	9. Relevancy	2.80	3.67	3.70	4.92	4.86	1.00	1.00
subtle	1. Coherence	4.16	4.72	4.52	4.80	4.89	3.00	3.42
	2. Cohesion	4.03	4.56	4.05	4.85	4.88	3.00	4.60
	3. Completeness	4.13	4.37	4.13	4.98	4.66	2.99	4.49
	4. Conciseness	2.51	2.07	2.29	3.44	4.77	2.22	2.20
	5. Correctness	4.35	4.76	4.31	4.50	4.92	3.00	3.38
	6. Informativeness	4.32	4.85	4.41	4.87	4.91	3.00	3.46
	7. Integration	4.08	4.49	3.98	4.93	4.90	3.00	4.87
	8. Readability	3.31	3.09	3.76	4.42	4.77	2.90	2.95
	9. Relevancy	4.31	4.63	4.50	5.00	4.85	2.98	3.48

Table 12: ORKGSynthesis dataset detailed evaluation results of various  $LLM_{eval}$  over  $LLM_{gen} = \text{LLaMA-3.1-70B}$ .

Set	Rubrics	M1	M2	M3	M4	M5	M6	M7
benign	1. Coherence	4.86	4.99	4.87	5.00	4.93	3.01	5.00
	2. Cohesion	4.86	4.98	4.78	4.99	4.90	3.00	5.00
	3. Completeness	4.47	4.65	4.62	4.76	4.70	3.00	4.93
	4. Conciseness	4.87	4.81	4.72	4.30	4.81	3.00	4.13
	5. Correctness	4.90	4.98	4.86	4.89	4.94	3.00	4.98
	6. Informativeness	4.78	4.98	4.91	4.97	4.93	3.06	5.00
	7. Integration	4.86	4.97	4.72	4.91	4.94	3.00	4.98
	8. Readability	4.89	4.97	4.90	4.44	4.87	3.00	4.98
	9. Relevancy	4.83	4.96	4.91	5.00	4.94	3.05	4.98
extreme	1. Coherence	2.74	3.75	3.54	4.47	4.91	1.04	1.00
	2. Cohesion	3.94	4.44	3.70	4.91	4.88	3.00	4.51
	3. Completeness	2.90	2.91	3.78	3.81	4.65	1.01	1.00
	4. Conciseness	1.81	1.31	1.90	2.87	4.69	1.01	1.00
	5. Correctness	3.84	4.15	4.01	4.31	4.92	1.01	1.00
	6. Informativeness	3.84	4.51	3.98	4.78	4.81	1.00	1.00
	7. Integration	3.82	4.34	3.60	4.86	4.89	2.89	3.52
	8. Readability	2.38	2.16	3.07	4.18	4.76	1.00	1.00
	9. Relevancy	3.06	3.70	3.83	4.98	4.88	1.00	1.00
subtle	1. Coherence	4.23	4.85	4.63	4.84	4.93	3.00	3.41
	2. Cohesion	4.16	4.61	4.11	4.96	4.92	3.00	4.83
	3. Completeness	4.40	4.50	4.30	4.99	4.73	3.00	4.79
	4. Conciseness	3.13	2.82	2.68	4.50	4.83	2.89	2.97
	5. Correctness	4.48	4.83	4.51	4.60	4.93	3.00	3.39
	6. Informativeness	4.46	4.90	4.70	4.91	4.87	3.04	3.49
	7. Integration	4.21	4.68	4.06	4.96	4.92	3.00	4.90
	8. Readability	3.71	3.21	3.86	4.53	4.82	2.91	2.98
	9. Relevancy	4.42	4.73	4.66	4.97	4.92	3.00	3.49

Table 13: ORKGSynthesis dataset detailed evaluation results of various  $LLM_{eval}$  over  $LLM_{gen} = \text{Qwen2.5-72B}$ .

Set	Rubrics	M1	M2	M3	M4	M5	M6	M7
benign	1. Coherence	4.79	4.95	4.83	5.00	4.92	3.00	4.98
	2. Cohesion	4.78	4.93	4.76	4.99	4.91	2.99	4.98
	3. Completeness	4.39	4.50	4.61	4.62	4.81	3.00	4.86
	4. Conciseness	4.82	4.80	4.71	4.26	4.75	2.98	4.07
	5. Correctness	4.87	4.97	4.87	4.89	4.95	3.00	4.93
	6. Informativeness	4.70	4.94	4.85	4.97	4.92	3.03	4.95
	7. Integration	4.77	4.92	4.75	4.90	4.95	3.00	4.97
	8. Readability	4.84	4.91	4.90	4.34	4.83	2.99	4.93
	9. Relevancy	4.70	4.92	4.83	5.00	4.94	3.01	4.90
extreme	1. Coherence	2.58	3.50	3.47	4.34	4.88	1.00	1.00
	2. Cohesion	3.72	4.29	3.62	4.87	4.90	2.99	4.60
	3. Completeness	2.76	2.76	3.61	3.67	4.58	1.00	1.00
	4. Conciseness	1.77	1.27	1.71	2.88	4.73	1.02	1.00
	5. Correctness	3.82	4.06	3.92	4.21	4.91	1.00	1.00
	6. Informativeness	3.69	4.43	3.90	4.83	4.87	1.00	1.00
	7. Integration	3.52	4.09	3.37	4.79	4.85	2.94	3.34
	8. Readability	2.36	2.10	2.97	3.84	4.62	1.01	1.00
	9. Relevancy	2.89	3.50	3.76	4.95	4.83	1.00	1.00
subtle	1. Coherence	4.17	4.76	4.45	4.78	4.97	3.00	3.33
	2. Cohesion	3.99	4.53	3.98	4.92	4.92	3.00	4.49
	3. Completeness	4.05	4.38	4.11	5.00	4.66	3.00	4.64
	4. Conciseness	2.98	2.66	2.60	4.22	4.85	2.90	2.89
	5. Correctness	4.36	4.78	4.45	4.50	4.92	3.00	3.35
	6. Informativeness	4.32	4.85	4.50	4.91	4.95	3.02	3.39
	7. Integration	4.05	4.50	4.03	4.90	4.89	3.00	4.91
	8. Readability	3.43	3.02	3.82	4.61	4.85	2.86	2.97
	9. Relevancy	4.30	4.68	4.52	4.99	4.96	3.00	3.42

Table 14: ORKGSynthesis dataset detailed evaluation results of various  $LLM_{eval}$  over  $LLM_{gen} = \text{Mistral-Large}$ .

	Set	$LLM_{gen}$	$LLM_{eval}$	Rubric	Rating	Rationale
1	subtle	LLaMA-3.1-8B	LLaMA-3.1-8B	Readability	4	The synthesis follows academic writing conventions almost perfectly and displays appropriate style. However, the last sentence is a bit unrelated to the research question and seems to be a non-sequitur, which slightly affects the overall rating.
2	benign	LLaMA-3.1-70B	SFT (benign)	Coherence	1	The synthesis provided does not form a coherent narrative, making it difficult to follow.
3	benign	Mistral-Large	SFT (benign) + RL (adversarial)	Integration	3	The synthesis integrates the sources into a coherent manner within a single paragraph. However, the transitions or connections between ideas could be smoother. For example, the transition from the information about Denmark to the information about England could be more fluid. Additionally, the synthesis could benefit from better markers of provenance/quotation to clearly indicate which study provided each piece of information.
4	benign	Qwen2.5-72B	SFT (benign) + RL (benign+adversarial)	Relevancy	5	The synthesis directly and consistently addresses the research question of how circRNAs are derived, providing a clear and concise overview of the back-splicing process and its key aspects.
5	adversarial	Mistral-Large	SFT (adversarial) + RL (benign+adversarial)	Accuracy	2	The synthesis contains factual inaccuracies regarding the regulatory mechanisms of circRNAs. While the overall structure is well-organized, some key claims about exon-skipping and alternative splicing are incorrect.
6	subtle	Qwen2.5-72B	LLaMA-3.1-8B	Coherence	4	The ideas are logically connected for the most part, but the narrative could be strengthened for better clarity. Some minor issues with sentence structure and wording were noted, but they do not significantly impact the overall coherence of the synthesis.
7	benign	LLaMA-3.1-70B	SFT (benign)	Informativeness	5	The synthesis is highly informative, providing valuable insights and detailed information that thoroughly addresses the research question.
8	subtle	LLaMA-3.1-70B	SFT (benign) + RL (adversarial)	Coherence	3	The ideas are logically connected for the most part, but the narrative could be strengthened for better clarity. The synthesis jumps abruptly from discussing the biogenesis of circRNAs to mentioning a mortality rate without adequately transitioning or providing context.
9	extreme	LLaMA-3.1-70B	SFT (benign) + RL (benign+adversarial)	Readability	1	The synthesis is poorly written, with pervasive issues in style, structure, and language use, making it difficult to understand. The sentence 'jets fans must be really confused with mark-sanchez performance as a 2nd string QB for philly!! Lmao' is entirely unrelated to the topic of eRNAs and disrupts the flow of the text.

Table 15: Example evaluation of LLM-generated responses based on different rubrics for the BioASQ dataset.



	Set	$LLM_{gen}$	$LLM_{eval}$	Rubric	Rating	Rationale
1	benign	LLaMA-3.1-70B	LLaMA-3.1-8B	Correctness	5	The synthesis accurately represents the content from the provided abstracts, without any factual errors or misinterpretations. The information is a faithful representation of the original sources, demonstrating a high level of correctness.
2	benign	Mistral-Large	SFT (benign)	Integration	5	The synthesis seamlessly integrates information from the various sources, using appropriate markers and connectors to create a smooth and unified narrative. All information is skillfully condensed into a single, well-structured paragraph.
3	extreme	Qwen2.5-72B	SFT (benign) + RL (adversarial)	Relevancy	1	The synthesis includes a relevant discussion about fake news detection methods, but it deviates significantly from the topic of detecting fake accounts. The last sentence about Stephen Curry, LeBron James, and Kevin Durant winning the 2024 Paris Olympics is completely out of scope and irrelevant to the research question.
4	extreme	Qwen2.5-72B	SFT (benign) + RL (benign+adversarial)	Completeness	1	The synthesis omits most of the relevant information, failing to capture the essential points or details from the provided abstracts. The last sentence of the synthesis is unrelated to the research question and appears to be a random statement about a sporting event.
5	subtle	LLaMA-3.1-70B	LLaMA-3.1-8B	Readability	5	The synthesis follows academic writing conventions perfectly and displays appropriate style, structure, and language use, making it exceptionally accessible and easy to read.
6	extreme	Qwen2.5-72B	SFT (benign)	Readability	5	The synthesis is exceptionally well-written, following stylistic and structural conventions with precise language use, making it accessible and easy to read. The text is well-structured and free of any noticeable issues with style or structure.
7	subtle	LLaMA-3.1-8B	SFT (benign) + RL (adversarial)	Conciseness	2	The synthesis includes some redundant or irrelevant statements, detracting from its clarity. For example, the synthesis repeats the phrase 'ultimately, finally, and in the end' and uses phrases like 'undoubtedly revealed, made evident, and clearly demonstrated' and 'highlighting, underscoring, and emphasizing the pressing, urgent, and critical need.' These repetitions make the synthesis less concise and more verbose than necessary.
8	extreme	Mistral-Large	SFT (benign) + RL (benign+adversarial)	Relevancy	1	The synthesis starts off well by addressing the research question, but it includes an unrelated sentence about Stephen Curry, LeBron James, and Kevin Durant, which is completely out of scope and lacks any connection to the topic of optimizing drug delivery methods to effectively cross the blood-brain barrier.

Table 16: Example evaluation of LLM-generated responses based on different rubrics for the ORKGSynthesis dataset.