Improving Biomedical QA Evaluation with Few Human Preference Signals

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) systems are increasingly used for biomedical Question Answering (QA). However, evaluating their effectiveness remains challenging due to the lack of reliable reference resources and the frequent misalignment between automatic evaluation metrics and human judgments. One practical question is how evaluation results should be used to guide development decisions, especially when they correlate weakly with human judgments and may even lead to different conclusions. We explore how to use a small amount of human preference signals to calibrate direct assessment evaluation metrics, and show that calibrated metrics consistently outperform baselines that either treat all metrics equally or rely on a single best metric across multiple datasets.

1 Introduction

Large Language Models (LLMs) (Luo et al., 2022; Bolton et al., 2024; Labrak et al., 2024; Kafkas et al., 2025) have shown promise in answering biomedical questions. Interest in Question Answering (QA) applications using Retrieval-Augmented Generation (RAG) techniques has also been on the rise (Lyu et al., 2024; Jin et al., 2024; Minbyul et al., 2024; Wang et al., 2024, for example). As a result, a recent direction has been to examine the use of RAG systems for discovery-oriented scientific questions that require long-form responses. A key problem, however, is the evaluation of their effectiveness, which remains underexplored.

Consider the question *"What proportion of Transcriptomic studies on plants found ICE1 Transcription factors to be associated with cold acclimation?"* as an example. Acquiring ground-truth answers for comparison is often infeasible due to the high cost of knowledge acquisition. In the absence of ground truth references, metric-based approaches to evaluation are a logical option. However, a naive use of metrics is not straightforward. When applying existing reference-free direct assessment evaluation metrics to evaluate domain-specific QA, e.g., measuring response faithfulness or entailment with relevant resources (Schimanski et al., 2024), previous studies have reported inconsistent findings regarding the effectiveness of individual metrics (Roychowdhury et al., 2024; Oliva et al., 2025). Furthermore, we highlight that metrics often lead to inconclusive assessments (§5.1) when applied to evaluate individual components of a RAG-based QA system. This raises the question of how to interpret such evaluation results to guide development decisions (e.g., should different metrics be weighted equally and combined, or should a single best metric be identified?).

Beyond problems with the application of metrics, Bavaresco et al. (2025) showed that domain experts apply stricter domain-specific criteria than automated metrics when evaluating LLM responses. Additionally, direct assessment itself is challenging for experts compared to pairwise comparison or reference-based assessment (Novikova et al., 2018). To this end, we work with biologists to collect human preferences for long-form answers from a RAG-based QA system and investigate how to calibrate the weights of direct assessment metrics from collected preferences. We find that a small amount of human preference signals is highly effective for calibrating weights, making the calibrated metrics substantially better than baselines with uniform or random weighting. In contrast, calibration based on LLM preferences, used as a low-cost proxy for human preferences, did not significantly improve alignment with human judgments.

Our contributions are: (1) a preference-guided framework for calibrating weights of direct assessment evaluation metrics using correlation analysis and logistic regression; (2) two human preference datasets FROSTTOLGENE and OZWHEATGENE, for RAG-based QA in plant and agricultural genomics, experiments on which (and an ex-

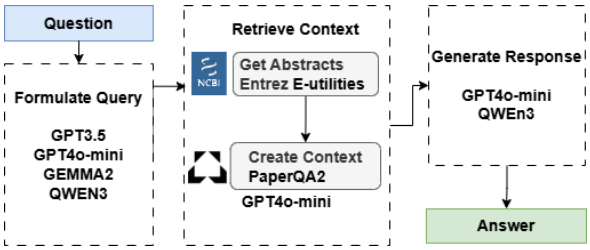

Figure 1: An overview of our RAG-based QA system.

isting SCIARENA benchmark) show that calibrated metrics align better with human judgments than uniform-weight or single-metric baselines; and (3) an exploration of LLM-based pairwise preference collection, as a cheap proxy, using generic and domain-specific rubric-based judges.

2 Problem Statement

We consider a 3-stage RAG-based system: (1) query formulation from user question; (2) retrieval and corresponding creation of content surrogates (i.e., summarizing retrieved documents) for inclusion in the prompt for answer generation; and (3) LLM answer generation. In an applied setting, decisions are needed about which variation is most effective for the proposed use case. Such decisions relate to the choice of LLM for each component or to algorithmic variations in any of the three above-mentioned steps.

Building on the RAGAS evaluation framework proposed by Es et al. (2024), we consider the scenario where ground-truth answers are unavailable, yet a practical evaluation approach is still needed to inform design decisions for the RAG-based QA system. Here, we first collect a pool of answers for a given question by running different variations of a QA system. Then, using existing automatic evaluation metrics, we rank the candidate answers and select a few candidates for human judgment. Finally, the collected human judgments and feedback are used to build better evaluation metrics and to inform design decisions for the QA system.

3 RAG-based QA System

To enable our study, we first built an in-house RAG QA system, GENERAG, since the existing biomedical QA systems are either not publicly available or focused solely on generating short answers. Figure 1 shows the pipeline of our system and design options for each stage.

Query Formulation. Our focus user group, consisting of plant genomics researchers, typically relies on NCBI's Entrez system[1] to search databases (e.g., PubMed) for relevant scientific information. To search via the Entrez API, we first convert a user question to search terms. To do this, we prompt LLMs with few-shot examples to: (1) generate a single query term for a given question (this setting is denoted as *SQ*); and, (2) create sub-questions from the given question and then generate a query term for each sub-question (denoted as *MQ*).

Context Retrieval. We retrieve relevant abstracts from PubMed focusing only on the open-access subset. A brief description of the API is provided in the App. §A. In the *MQ* setting, where multiple query terms are generated, two search variants are implemented in GENERAG: (1) concatenating all query terms into a single query string using the `OR` operator (MQ_{+Cat}); and (2) retrieving abstracts separately for each query term and concatenating them (MQ_{+Stack}). After retrieving the abstracts, we follow PaperQA2 (Skarlinski et al., 2024) pipeline, using *GPT4o-mini* to chunk, filter, re-rank, and summarize them. The final context is a set of summarized relevant excerpts from the abstracts.

Answer Generation. Given the user's question and the context from the previous step, the system prompts an LLM to generate the final answer.

Models Selection. In the *SQ* setting, we use four LLMs: *GPT3.5-turbo*, *GPT4o-mini*, *GEMMA2-9B*, and *Mistral-Nemo-12B*. In the *MQ* setting, we use *GPT4o-mini* and *Mistral-Nemo-12B*. For answer generation, we use *GPT4o-mini* and *QWEN-3-8B*. We evaluated several domain-specific (BioMistral-7B, MedLLAMA-7B, PMC-LLAMA-7B and BioGPT-Large) and generic LLMs (QWEN2.5-32B and Claude-3-Sonnet) on a question subset, selecting models via a quality control inspection.

4 Datasets

In collaboration with two plant genomics scientists, we compiled two new datasets. The scientists selected candidate wheat genes identified through their own genomic annotation research and posed questions regarding the association of these genes with specific traits. The resulting questions are divided into two sets (examples in App. Table 11):

[1] https://www.ncbi.nlm.nih.gov/search/

- FROSTTOLGENE, a set of 42 questions about the relevance of genes to frost tolerance and cold acclimation, using genes identified through a literature search by the domain experts.

- OZWHEATGENE, a set of 23 questions from genome-wide association analysis that identified genes linked to crop yield and component traits for crop yield for the genes from *Ozwheat* diversity panel (Hyles et al., 2024). These questions are related to multiple genes and physical traits of wheat.

4.1 Collecting Human Preferences

We collected expert preferences over pairs of answers from the FROSTTOLGENE and OZWHEATGENE datasets through an internally developed survey application (details in App. §H). For each question, we generated a pool of 12 answers using different variants (i.e., different combinations of RAG system parameters) using our QA system.

In cases where the answer space is combinatorial, it is common to sample and choose candidate LLM answers for human annotations, given the complexity and time requirement. For example, Feng et al. (2025) selected the top 10 candidates via maximum-discrepancy competition and WildBench (Lin et al., 2025) discarded highly similar chats from evaluation. Furthermore, Pan et al. (2026) showed that in the Direct Preference Optimization (DPO) context, the human preference data should cover both good and bad responses. Considering the substantial human annotation effort required to judge pairwise answers with large retrieved contexts, we also sampled candidate answers from the answer pool (66 unique combinations of answer pairs per question). Given our focus on calibrating metric weights for only a specific set of metrics, we used the aggregated evaluation scores as the basis for selecting three candidate answers per question, corresponding to the highest-, moderate-, and lowest-quality answers.

We adopted a tournament-style sampling strategy, where the winner from the first random pair moves forward and is paired with the remaining third answer, as we found when comparing annotations on 'all' pairs and fewer pairs in tournament-style setup (paired randomly), they both achieve close inter-annotator agreements Cohen's κ of 0.68 and 0.64, respectively. Critically, the annotators were not aware of the quality ranks of the answers. During the annotation process, the annotators could only see the generated contexts and the answers, and did not have access to the formulated query or the retrieved set of abstracts. In addition to selecting their preferred answer, the annotators were instructed to provide a short explanation of the problems they identified in the unpreferred answer (referred to later as the *human reasonings*).

Of the 65 questions across both datasets, two questions (one from each dataset) were reserved for annotator training and excluded from analysis. For the remaining 63 questions, we collected annotations on answer pairs, resulting in 166 pairwise comparisons. Three domain expert annotators annotated ten questions, and the remaining 53 were single-annotated. For the multi-annotations, we further considered the mode over the multi-annotations and concatenated the reasons. The final data set totalled 130 unique annotations.

4.2 Preferences by LLM-as-a-Judge

To compensate for the high cost of human expert annotations, which invariably limits the number of collected human annotations, we developed a rubric-based LLM-as-a-judge baseline (Hashemi et al., 2024; Lin et al., 2025; Zhao et al., 2025) and constructed a synthetic dataset containing LLM preferences over all pairwise answers. Following Lin et al. (2025), we first created a set of generic rubric questions and a set of domain-specific rubric questions in consultation with domain experts. We then prompted a *GPT-4o-mini* model to give preferences over pairwise answers using the rubric as a guideline. We first collected LLM preferences over the human-annotated subset under four settings: LLMDIRECT—direct LLM preference without rubrics; LLMGENERIC—rubrics containing only generic questions; LLMDOMAIN—rubrics containing only domain-specific questions; and LLMALL—rubrics containing both generic and domain-specific questions. The prompts are included in App. §E. After generating the preferences, we measured human–LLM agreement using Cohen's κ and correlation using Pearson's r. As shown in Table 2, the LLMDOMAIN achieved the highest agreement and correlation with human preferences. Therefore, we used the LLMDOMAIN prompts to collect 4,290 LLM preferences over all answer pairs for each question.

Quality Dimension	Metric	Evaluation Dimension
Semantic Similarity	ROUGE-1 (Lin, 2004), BERTScore (Zhang et al., 2020), CosSim. (Reimers and Gurevych, 2019)	$Qr2Q$, $C2Ab$, $A2C$
Relevance (LLM-based)	Context Relevance (Es et al., 2024)	$C2Q$
	Answer Relevance (Es et al., 2024)	$A2Q$
Factual consistency	FactCC (Kryscinski et al., 2020)	$C2Ab$, $A2C$
	AlignScore (Zha et al., 2023)	$Qr2Q$, $C2Q$, $A2Q$, $C2Ab$, $A2C$
Faithfulness	HHEM-2.1-open (Es et al., 2024)	$C2Ab$, $A2C$

Table 1: Evaluation metrics selected for different evaluation dimensions. $Qr2Q$ =query-to-question, $C2Q$ =context-to-question, $A2Q$ =answer-to-question, $C2Ab$ =context-to-abstract and $A2C$ =Answer-to-Context.

Prompt Strategy	Cohen's κ	Pearson's r
LLMDIRECT	0.46	0.409
LLMDOMAIN	**0.55**	**0.530**
LLMGENERIC	0.50	0.486
LLMALL	0.53	0.515

Table 2: Agreement between human vs. LLM preferences in four LLM-as-a-Judge settings (p-val<.001).

4.3 Inter-annotator Agreements

We computed agreement on the common-answer pairs shared across annotators. The overall agreement level is moderate with κ =0.68. Additionally, we also calculated the Leave-One-Out (LOO) inter-annotator agreement scores to assess individual annotators' impact on the agreement level. The LOO Cohen's κ scores were 0.31 for *annotator 1*, 0.12 for *annotator 2*, and 0.32 for *annotator 3*, indicating that the annotators have higher variability in individual alignment with the remaining annotators. This also highlights the challenge of conducting human evaluation in highly specialized domains, since domain experts usually have slightly different background knowledge, which can strongly influence their judgments.

5 Automatic Evaluation vs. Preferences

Drawing from existing evaluation frameworks in the literature, we first identified a set of quality dimensions and corresponding metrics for the evaluation of RAG-based systems. Beyond assessing only the quality of final answers, following Es et al. (2024), we used these metrics to assess individual components of the GENERAG pipeline.

5.1 Metrics and Dimensions

We applied multiple metrics across five evaluation dimensions. Specifically, we assessed the quality of the formulated query, the retrieved context, and the final answers to the user questions, which we refer to as in the $Qr2Q$, $A2Q$ and $C2Q$ evaluation, respectively. Additionally, in $A2C$ evaluation, the generated answers were checked against the context, and in $C2Ab$ evaluation, the generated contexts by PaperQA2 were evaluated against the retrieved abstracts. We selected the semantic similarity, relevance, factual consistency and faithfulness quality dimensions and adopted suitable metrics for each dimension. The metric list is presented in Table 1. Altogether, we considered 18 evaluation metrics. (Details in App. §C.)

Setting	Query LLM	Ans LLM	$Qr2Q$ Avg.	$C2Q$ Avg.	$C2Ab$ Avg.	$A2Q$ Avg.	$A2C$ Avg.
SQ	GPT3.5	GPT4	<u>0.606</u>	0.154	0.465	0.793	0.408
	GPT3.5	QWEN				0.813	**0.497**
	GPT4	GPT4	0.596	0.175	0.486	0.789	0.414
	GPT4	QWEN				**0.823**	<u>0.484</u>
	GEMMA	GPT4	0.596	0.183	0.471	0.774	0.422
	GEMMA	QWEN				0.804	0.478
	Mistral	GPT4	0.567	0.220	<u>0.481</u>	0.792	0.391
	Mistral	QWEN				0.811	0.473
MQ_{+Cat}	GPT4	GPT4	**0.607**	0.264	**0.521**	0.788	0.392
	GPT4	QWEN				0.779	0.399
	Mistral	GPT4	0.564	0.268	0.517	0.807	0.421
	Mistral	QWEN				0.785	0.433
MQ_{+Stack}	GPT4	GPT4	0.579	**0.313**	0.511	0.804	0.399
	GPT4	QWEN				0.748	0.423
	Mistral	GPT4	0.578	<u>0.312</u>	0.511	<u>0.812</u>	0.399
	Mistral	QWEN				0.743	0.422

Table 3: Metric scores for FROSTTOLGENE query, context and answer evaluation in three settings: *SQ*, MQ_{+Cat} and MQ_{+Stack} (Best **bolded**, and second-best <u>underlined</u>).

From the metric scores, we found that none of the variants *alone* is consistently performing well across different evaluation dimensions, and it is not possible to conclusively decide which LLMs and design options to use in the various stages of the RAG system based on them. For example, summarized results on FROSTTOLGENE in Table 3 shows, the best context was obtained by GPT4 in the MQ_{+Stack} setting for $C2Q$, but for $C2Ab$, MQ_{+Cat} setting obtained the highest results. More results and discussion are in the App. §D. This raised our next question: if these metrics cannot reach consensus, how can we use them to guide design decisions for the QA system?

5.2 Metric vs. Preferences

We started by analysing the alignment of the human preferences and aggregated metric scores on FROSTTOLGENE and OZWHEATGENE datasets. We converted the scores and the human into binary preference labels. Formally, for each question q

and answer pair (a, b), a label of 0 is assigned if annotators prefer a, and 1 if they prefer b. Metric-based preferences are derived analogously:

$$Pref(a, b|\mathcal{S}) = \begin{cases} 0, & \text{if } \mathcal{S}(a) \geq \mathcal{S}(b), \\ 1, & \text{otherwise.} \end{cases} \quad (1)$$

where uniform weight $\mathcal{S}(a_i) = \frac{1}{n}\sum_{j=1}^{n} M_j(a_i)$ calculated over $M_j \in \mathcal{M}$ metrics. Using these labels, we computed the rate of cases when human annotators selected metric-based preferred answers. All metrics were considered equally, equivalent to an **uniform-weighting**. The resulting agreement percentages and Pearson's correlation coefficients r, along with the statistical significance reported in App. Table 17. Overall, experts agreed with the metric-based preferences in 61.45% of the cases (Kappa κ = 0.30). We also measured a moderate but significant Pearson correlation coefficient, r = 0.300 (p < 0.01) between human-metric preferences.

Further, between human and individual metrics-based preferences, we found that the agreement rates vary substantially across metrics. We found *A2C-AlignScore* has the highest human-metric preference agreement of 65.75%. Additionally, *A2Q-AlignScore* achieved 61.64% agreement, *C2Ab-FactCC* 61.45% and *A2C-Faithfulness* achieved 61.64% human-metric preference agreements.

In light of the findings, we infer that not all automatic metrics correlate uniformly with human preferences, and it is not possible to have a conclusive decision on the metrics to choose. Thus, in the next section, we explore ways to incorporate human preferences into metric reliability calibration.

6 Preference-Guided Weight Calibration

Before applying or aggregating evaluation metrics, we want to determine which metrics align better with human preferences. The human–metric preference agreement rates computed using uniformly weighted metrics in the previous section allow us to identify which metrics contribute more strongly to alignment with human judgments. These agreements can be considered as a proxy for human preference alignments.

Based on this, we propose a preference-guided approach for calibrating metric reliability weights using human preferences. Formally, let the agreement rate between human and metric preferences for metric M_j be denoted as the calibration weight w_j. Using these weights, the adjusted aggregated score for an answer a_i is computed as $\mathcal{S}_{adj}(a_i) = \frac{1}{n}\sum_{j=1}^{n} w_j \times M_j(a_i)$. The corresponding preference labels, $Pref(a, b \mid \mathcal{S}_{adj})$, are then obtained using Eq. 1. We consider two approaches to calibrate the metric weights—directly using human–metric preference agreement as weights (DMW), and learning the weights using a logistic regression model (LMW).

Direct Metric Weighting (DMW) We considered human-metric preference agreements directly as the metric weights. By dividing the dataset into a calibration and validation set and computing human–metric preference agreement scores for each metric on the calibration set, we obtained the calibrated weights for the metrics. We then applied these weights to the validation set for aggregating the metric scores and obtaining the weighted-metric preferences, $Met_{dir+hum}$. Finally, we measured the human-$Met_{dir+hum}$ agreement rate and compared it with the agreement rates between human vs. uniform- and random-weighted metric preferences, Met_{unif} and Met_{rand}, respectively. Similarly, we calibrated weights using LLM preferences and obtained metric preferences $Met_{dir+llm}$ over the human validation sets and measured the agreement rates against human preferences.

Learned Metric Weighting (LMW) To learn the metric weights from the human preferences, we built separate logistic regression models for each metric. For each metric, we prepared the dataset with pairwise high and low metric scores (implying metric-based preference) for each answer pair, and human preference [0,1] as the label. The model's prediction task was essentially a binary classification task to predict human preference for a pair of high-low metric scores. We then considered the training accuracy of the model as the metric weight. The learned metric weights were then applied to the human validation subsets to obtain $Met_{log+llm}$ and validated against the human preferences.

6.1 Experimental Setup

We randomly split the human-annotated dataset into a 60%-40% calibration–validation split, yielding 78 calibration and 52 validation instances per set, repeating this process 100 times to report the mean and standard deviation of the agreement rates of the human-metric preferences. For experiments involving LLM-generated preferences, we randomly sampled a larger calibration set by

No Calibration		
Uniform weighting	61.44 ± 5.95	
Random weighting	55.42 ± 7.21	
W/ Calibration		
	DMW	LMW
Human pref.	64.30 ± 4.34	70.00 ± 4.05
LLM pref.	53.17 ± 2.45	62.95 ± 4.10

Table 4: Agreement rate (%) and std. dev. (CI=95%) between human and metric-based preferences, computed via Direct Metric Weighting (DMW) and Learned Metric Weighting (LMW), and compared against random and uniform weights on the validation set.

Eval. Dimension	Uniform	DMW	LMW
All	61.44	64.30	70.00
$A2C$-AlignScore (best)	**66.03**	**64.19**	**66.04**
$Qr2Q$	50.92	**52.88**	**52.17**
$C2Q$	**52.84**	52.85	51.88
$C2Ab$	45.07	48.54	47.64
$C2Q$ +$C2Ab$	48.57	49.57	48.66
$A2Q$	48.19	50.80	46.58
$A2C$	**61.15**	**62.03**	**58.85**
$A2Q$ +$A2C$	**58.54**	**60.35**	**61.55**

Table 5: Human-metric preference agreements(%) for the calibration using subsets of metrics compared to 'All' metrics. The top 4 best-performing subsets are bolded for each setting.

taking 80% of the annotations to derive the calibrated weights for $k = 100$ iterations in both DMW and LMW settings. For DMW, we averaged the 100 calibration weights to generate a single set of weights. Validation was done on the 100 human-annotated validation folds described above. We also considered two baselines: uniform weighting and random weighting, where all metrics are assigned equal weights or random weights, respectively.

6.2 Results and Discussion

Calibration with Human Preferences. Human preference-based calibration leads to a significant improvement in alignment between human and metric preferences on the validation set, with an increase in the average agreement rate from 61.44% with uniform metric-weights to 64.30% with DMW and 70.00% with LMW methods (Table 4). This result is also significantly higher than the *random-weights* baseline with an agreement level of 55.42%.

We observed a moderate Pearson's correlation coefficient $r = 0.315$ for $Met_{dir+hum}$-human and $r = 0.408$ for $Met_{log+hum}$-human preference labels, compared to a lower correlation of $r = 0.281$ for Met_{unif}-human labels across all k validation samples (for all, p<0.0001). Furthermore, we found that the improvement in the human-metric agreement rates is also statistically significant. With a one-tailed (Group A > Group B) Wilcoxon's signed-rank test (Riffenburgh, 2006), we measure the agreement rates using DMW vs. uniform-weighting over k samples and observed a statistically significant improvement $z = 2.865$ (p<0.01, significant at $\alpha = 0.005$ level) and which is for LMW, $z = 0.178$ (p<0.01, significant at $\alpha = 0.05$ level), implying a positive significance with a tighter bound of standard deviations.

Comparison of Best Metrics and Subsets. To decide on which evaluation dimensions were most effective in weight calibration, we use $Qr2Q$, $C2Ab$, $C2Q$ +$C2Ab$, $A2Q$ +$A2C$ subsets separately and the best metric, $A2C$-AlignScore (based on App. Table 17) to calibrate weights. The results are reported in Table 5. Under the DMW setting, the single best metric achieved 64.19%, which was close to the 64.30% obtained when using all metrics, implying that under direct metric calibration, considering one best metric is sufficient. In contrast, under the LMW setting, neither the best-performing metric nor any individual subset of metrics achieved a score close to the 70% obtained using metrics from all dimensions. Furthermore, the $Qr2Q$, $A2Q$, and $A2C$ metrics appear to contribute the most toward learning the metric weights.

Calibration with LLM-preferences. We found that, in the absence of human-annotated data, LLM-generated preferences can also be used to calibrate metric weights. As shown in Table 4, the DMW setting achieves only 53.17% agreement with human judgments, which is lower than both uniform and random-weighted baselines. In contrast, the LMW setting improves this agreement to 62.95%. Although calibration using LLM preferences does not reach the 70% agreement achieved with human-preference-based calibration, it provides a better alternative to uniform metric aggregation when human annotations are unavailable. Across all k samples, we also observe a Pearson correlation of $r = 0.293$ (p < 0.001) between human-$Met_{dir+llm}$ labels, and $r = 0.396$ (p < 0.001) between human and $Met_{log+llm}$ labels, further indicating that LLM-preference in LMW is relatively closer to human-preference-based calibration.

	Unif. Weighting	DMW	LMW
Combined	53.39	59.82	**60.51**
Biomed.	52.14	**60.17**	59.82
Medicine	53.91	**58.77**	58.75
Neuroscience	54.00	61.09	**62.18**
Pharmacy	52.42	60.05	**60.57**
Public Health	51.77	59.00	**61.23**

Table 6: SCIARENA human-metric preference agreements(%).

Evaluation on the SciArena Benchmark. To test the generalizability of the calibration techniques on an out-of-domain dataset, we conducted experiments on SCIARENA (Zhao et al., 2025), which is a benchmarking dataset with annotations collected from the researcher community over pairwise LLM-generated responses on scientific literature tasks. We considered *'Biomedical Engineering'*, *'Medicine'*, *'Neuroscience'*, *'Pharmacy'* and *'Public Health'* subsets, as the domains were close to our focus. Details of the data size of each subset are reported in the App. §G, and an example QA pair is in Table 25. We applied both DMW and LMW techniques on each subset separately and combined. Results are in Table 6. We observe that in all cases human-metric agreement rates improved from the uniform weighting, and the best agreement rate of 60.51%, on the combined set, was achieved in the LMW setting. Note, in the SciArena dataset, only answers and context (citations) are available, and we used only 9 metrics covering $C2Q$,$A2Q$ and $A2C$. As shown in §6.2, the LMW setting performed better with a larger set of metrics, particularly with $Qr2Q$, $A2Q$, and $A2C$, which might be a possible reason for the improvement in agreement from DMW to LMW.

Quality Dimensions from Expert Reasons. To understand which quality aspects the experts prioritize, we prompted a *GPT4o-mini* model with a one-shot example to group the human rationals and assign interpretable labels (referred to as *criteria*). In case multiple reasons were present in one rationale, we split and consider them separately. Using another prompt by adding these dimensions in-prompt, we asked the LLM to assign a label to each of the human reasons, which were further corrected through manual checking. The prompt and resulting criteria are in App. §F. We found that the main factors influencing human judgments are answer and context relevance, lack of detail in the answer and incorrect or irrelevant context (App. Table 22). This validates our contextual appraisal of quality based on a set of component-wise metrics for RAG systems.

We also observed 35 cases where annotators noted that the answers were very similar, resembling a 'tie' between two answers and that both answers are good/bad (Chiang et al., 2024). Two such examples are in App. Table 23.

In addition, human experts also considered answer clarity, fluency, organization, and structure. This finding resonates with prior work that incorporates stylistic qualities (D'Souza et al., 2025) and answer usefulness (Malaviya et al., 2024) in LLM-based evaluation.

We further analyzed cases where metric-based preferences differ from human judgments (App. Table 21), highlighting the metrics' inability to cover certain evaluation criteria. For example, when only $C2Ab$ metrics are used to derive metric preferences, they disagree with 30 human preferences in cases where humans based their judgments on 'Relevance of Answer'. The lowest amount of mismatching was obtained when all metrics were considered to generate preferences.

Finally, we observed that the qualitative aspects of the answers can be a matter of subjective judgments, varying across expertise levels. App. Table 24 gives an example of such an annotation, where both annotators considered different rationales for the same preference. Therefore, effectively selecting an evaluation strategy for RAG systems requires incorporating metrics from all components of the RAG pipeline and human-in-the-loop evaluation strategies.

7 Related Work

We first explored the limitations of different QA datasets and language models in the biomedical domain. Then, we discussed challenges with the existing automated metrics for RAG-based QA systems and how human preferences are integrated to improve LLM-based evaluations.

Biomedical QA Datasets. Several general and specialized datasets support biomedical QA. MedQA (Jin et al., 2021) and MedRedQA (Nguyen et al., 2023) targeted general medical knowledge, whereas GeneTuring (Shang et al., 2025) introduced gene-centric questions, PubMedQA (Jin et al., 2019) and BIOASQ-QA (Krithara et al., 2023) focused on literature-grounded questions. More recently, ExpertQA (Malaviya et al., 2024) introduced expert-curated questions across multiple

domains. However, these datasets are not suitable for evaluating long-form, open-ended responses. Our study introduced qualitative aspects– such as relevance and alignment with the question and supporting context–through calibrated metrics by human preferences.

LLMs for Answering Domain-specific Questions. LLMs have been pretrained or fine-tuned on biomedical corpora to support domain-specific QA (Bolton et al., 2024; Luo et al., 2022; Wu et al., 2024; Labrak et al., 2024; Yasunaga et al., 2022). Some models focused on gene-related tasks, such as Geneverse (Liu et al., 2024) and GP-GPT (Lyu et al., 2024). In parallel, RAG has been explored for biomedical QA using literature retrieval (Skarlinski et al., 2024; Patel et al., 2026), retrieval-based gap-analysis and knowledge assertion (Ning et al., 2025) or combining databases with external tools (Wang et al., 2024). However, such systems rely on restricted resources, limiting public availability and reproducibility.

Evaluation of RAG QA systems. Evaluating RAG-based QA models remain challenging due to the lack of reliable reference answers. RAGAS (Es et al., 2024) proposed a reference-free evaluation framework judging faithfulness, relevance, and groundedness. Adlakha et al. (2024) evaluated correctness and faithfulness in instruction-following QA and Zha et al. (2023) evaluated factual consistency via AlignScore. Benchmarks such as LitQA2 (Skarlinski et al., 2024) and LabBench (Laurent et al., 2024) assessed QA accuracy and biological reasoning skills, while ExpertQA (Malaviya et al., 2024) evaluated qualitative dimensions including usefulness and cite-worthiness. However, guidance on metric selection for domain-specific QA remains limited.

Human-preferences in LLM Evaluators. LLMs were used as automated evaluators to judge AI-generated responses (Li et al., 2025; Hijazi et al., 2024; D'Souza et al., 2025). However, there are significant gaps between coverage of human-centred criteria by LLMs (Miller and Tang, 2025) or inconsistent correlations with human judgment (Oliva et al., 2025). Consequently, several approaches now integrate human feedback into LLM-evaluators for automated evaluation of reasoning chains (Jacovi et al., 2024) and to align automated evaluation with human preferences (Yue et al., 2025; Polo et al., 2025; Ye et al., 2025).

8 Conclusions

Evaluating the effectiveness of RAG-based QA systems for generating long-form answers to biomedical questions remains a challenging task, regardless of the availability of several automatic evaluation metrics. Our analysis revealed substantial variability across metrics when evaluating different components of a RAG system. Such variability makes it difficult to interpret results, select appropriate metrics, or identify a single reliable metric for evaluation. Furthermore, we found that no individual metric consistently aligns with expert judgments.

To address this issue, we considered the human–metric preference agreement rates as indicators of each metric's contribution toward human alignment and proposed calibrating metric weights before applying or aggregating them. Our results showed that, without calibration, relying on the single best-performing metric can achieve performance comparable to direct metric weighting (DMW). However, the strongest alignment with human preferences was achieved when metrics across all evaluation dimensions—query, context, and answer—were considered and the weights were learned using the LMW approach.

We further conducted similar weight calibration using LLM preferences collected through a rubric-based LLM-as-a-judge baseline. The results showed that only the learning-based LMW approach improved human alignment of the metrics over uniform weighting. However, the performance remained well below the human-preference-based calibration.

Additionally, to conduct the weight calibrations, we create two expert-curated plant genomics QA datasets that address the scarcity of non-human genetics QA benchmarks, and extended the analysis to the SCIARENA benchmark dataset.

Future work includes extending this analysis to other domains or task types to test the generality of preference-guided metric calibration and adapting other learning techniques for improved metric weight calibration. In addition, integrating preference-guided metrics into training objectives for RAG-based QA models—through reinforcement learning or self-refinement—could further improve alignment with expert human judgments.

9 Limitations

As a proof of concept, we propose using a small amount of human preference signals to calibrate

existing direct assessment metrics. We select representative metrics that aim to cover different quality dimensions. We do not claim that our choice of metrics is exhaustive. Moreover, given the complexity of discovery-oriented scientific questions and long-form answers, these expert annotations are inherently subjective and challenging to scale consistently, which constitutes a key limitation.

10 Ethics Statement

This work has received ethics approval for human annotation collection survey through the institutional ethics approval process. We will provide full details in the accepted, camera-ready version.

References

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, 12:681–699.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, et al. 2024. "biomedlm: A 2.7b parameter language model trained on biomedical text". *arXiv preprint arXiv:2403.18421*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 8359–8388. PMLR.

Jennifer D'Souza, Hamed Babaei Giglou, and Quentin Münch. 2025. YESciEval: Robust LLM-as-a-judge for scientific question answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 13749–13783, Vienna, Austria. Association for Computational Linguistics.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158.

Kehua Feng, Keyan Ding, Tan Hongzhi, Kede Ma, Zhihua Wang, Shuangquan Guo, Cheng Yuzhou, Ge Sun, Guozhou Zheng, Qiang Zhang, and Huajun Chen. 2025. Sample-efficient human evaluation of large language models via maximum discrepancy competition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 10913–10947, Vienna, Austria. Association for Computational Linguistics.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. LLM-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 13806–13834, Bangkok, Thailand. Association for Computational Linguistics.

Hashem Hijazi, Diego Molla, Vincent Nguyen, and Sarvnaz Karimi. 2024. Using large language models to evaluate biomedical query-focused summarisation. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 236–242.

Jessica E. Hyles, Howard A. Eagles, Kerrie Ramm, Bjorg Sherman, Andrew Gock, Sandra Stops, Tanya Phongkham, Emmett Leyne, Tina Rathjen, Radoslaw Suchecki, Lauren Stevens, Louise Ord, Nick S. Fradgley, Meredith D. McNeil, Annelie Marquardt, Samuel C. Andrew, Kerrie Forrest, Russell F. Eastwood, Adam Norman, Annette Tredrea, Richard Trethowan, Ben Trevaskis, and Shannon K. Dillon. 2024. Ozwheat: a genome-to-phenome platform to resolve complex traits for wheat pre-breeding and research. *bioRxiv*.

Alon Jacovi, Yonatan Bitton, Bernd Bohnet, Jonathan Herzig, Or Honovich, Michael Tseng, Michael Collins, Roee Aharoni, and Mor Geva. 2024. A chain-of-thought is as strong as its weakest link: A benchmark for verifiers of reasoning chains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 4615–4634, Bangkok, Thailand. Association for Computational Linguistics.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14).

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th In-*

ternational Joint Conference on Natural Language Processing, EMNLP-IJCNLP, pages 2567–2577.

Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. 2024. Genegpt: augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics*, 40(2).

Şenay Kafkas, Marwa Abdelhakim, Azza Althagafi, Sumyyah Toonsi, Malak Alghamdi, Paul N. Schofield, and Robert Hoehndorf. 2025. The application of large language models to the phenotype-based prioritization of causative genes in rare disease patients. *Scientific Reports*, 15(1).

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. BioASQ-QA: A manually curated corpus for Biomedical Question Answering. *Scientific Data*, 10(1).

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 9332–9346, Online. Association for Computational Linguistics.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics*, pages 5848–5864.

Jon M Laurent, Joseph D Janizek, Michael Ruzo, Michaela M Hinks, Michael J Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D White, and Samuel G Rodriques. 2024. Lab-bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362*.

Qintong Li, Leyang Cui, Lingpeng Kong, and Wei Bi. 2025. Exploring the reliability of large language models as customized evaluators for diverse NLP tasks. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10325–10344.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. 2025. Wildbench: Benchmarking llms with challenging tasks from real users in the wild. In *International Conference on Learning Representations*, volume 2025, pages 47852–47870.

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

Tianyu Liu, Yijia Xiao, Xiao Luo, Hua Xu, Wenjin Zheng, and Hongyu Zhao. 2024. Geneverse: A collection of open-source multimodal large language models for genomic and proteomic research. In *Findings of the Association for Computational Linguistics, EMNLP*, pages 4819–4836.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings Bioinform.*, 23(6).

Yanjun Lyu, Zihao Wu, Lu Zhang, Jing Zhang, Yiwei Li, Wei Ruan, Zhengliang Liu, Xiaowei Yu, Chao Cao, Tong Chen, et al. 2024. Gp-gpt: Large language model for gene-phenotype mapping. *arXiv preprint arXiv:2409.09825*.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. ExpertQA: Expert-curated questions and attributed answers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3025–3045.

Justin K Miller and Wenjia Tang. 2025. Evaluating llm metrics through real-world capabilities. *arXiv preprint arXiv:2505.08253*.

Jeong Minbyul, Sohn Jiwoong, Sung Mujeen, and Kang Jaewoo. 2024. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*, 40.

Vincent Nguyen, Sarvnaz Karimi, Maciej Rybinski, and Zhenchang Xing. 2023. MedRedQA for medical consumer question answering: Dataset, tasks, and neural baselines. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 629–648.

Yingpeng Ning, Yuanyuan Sun, Ling Luo, Yanhua Wang, Yuchen Pan, and Hongfei Lin. 2025. Medtrust-rag: Evidence verification and trust alignment for biomedical question answering. In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1791–1794.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 72–78, New Orleans, Louisiana. Association for Computational Linguistics.

Maria Paz Oliva, Adriana D. Correia, Ivan Vankov, and Viktor Botev. 2025. The illusion of a perfect metric: Why evaluating AI'S words is harder than it looks. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 832–842, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

James M. Ostell. 2012. Entrez: The ncbi search and discovery engine. In *Data Integration in the Life Sciences*, pages 1–4.

Yu Pan, Zhongze Cai, Huaiyang Zhong, Guanting Chen, and Chonghuan Wang. 2026. What matters in data for dpo? *Advances in Neural Information Processing Systems*, 38.

Vatsal Pravinbhai Patel, Elena Jolkver, and Anne Schwerk. 2026. MedDiscover: A domain-specific retrieval-augmented generation framework for evidence-grounded knowledge extraction in metabolomics. *Comput. Struct. Biotechnol. J.*, 35(1).

Felipe Maia Polo, Xinhe Wang, Mikhail Yurochkin, Gongjun Xu, Moulinath Banerjee, and Yuekai Sun. 2025. Bridging human and LLM judgments: Understanding and narrowing the gap. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China.

Robert H. Riffenburgh. 2006. Chapter 6 - statistical testing, risks, and odds in medical decisions. In *Statistics in Medicine*, second edition, pages 93–114. Academic Press, Burlington.

Sujoy Roychowdhury, Sumit Soman, HG Ranjani, Neeraj Gunda, Vansh Chhabra, and Sai Krishna Bala. 2024. Evaluation of rag metrics for question answering in the telecom domain. In *ICML Workshop on Foundation Models in the Wild*.

Tobias Schimanski, Jingwei Ni, Mathias Kraus, Elliott Ash, and Markus Leippold. 2024. Towards faithful and robust LLM specialists for evidence-based question-answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 1913–1931.

Xinyi Shang, Xu Liao, Zhicheng Ji, and Wenpin Hou. 2025. Benchmarking large language models for genomic knowledge with geneturing. *Briefings in Bioinformatics*, 26(5).

Michael D Skarlinski, Sam Cox, Jon M Laurent, James D Braza, Michaela Hinks, Michael J Hammerling, Manvitha Ponnapati, Samuel G Rodriques, and Andrew D White. 2024. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*.

Tatiana Tatusova. 2010. Genomic databases and resources at the national center for biotechnology information. In *Data Mining Techniques for the Life Sciences*, pages 17–44.

Chengrui Wang, Qingqing Long, Meng Xiao, Xunxin Cai, Chengjun Wu, Zhen Meng, Xuezhi Wang, and Yuanchun Zhou. 2024. Biorag: A rag-llm framework for biological question reasoning. *arXiv preprint arXiv:2408.01107*.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9).

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 8003–8016.

Ziyi Ye, Xiangsheng Li, Qiuchi Li, Qingyao Ai, Yujia Zhou, Wei Shen, Dong Yan, and Yiqun Liu. 2025. Learning llm-as-a-judge for preference alignment. In *The Thirteenth International Conference on Learning Representations, ICLR*.

Tan Yue, Rui Mao, Xuzhao Shi, Shuo Zhan, Zuhao Yang, and Dongyan Zhao. 2025. QAEval: Mixture of evaluators for question-answering task evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 14717–14730.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR*.

Yilun Zhao, Kaiyan Zhang, Tiansheng Hu, Sihong Wu, Ronan Le Bras, Taira Anderson, Jonathan Bragg, Joseph Chee Chang, Jesse Dodge, Matt Latzke, Yixin Liu, Charles McGrady, Xiangru Tang, Zihang Wang, Chen Zhao, Hannaneh Hajishirzi, Doug Downey, and Arman Cohan. 2025. Sciarena: An open evaluation platform for foundation models in scientific literature tasks. In *Proceedings of the NeurIPS 2025 Datasets & Benchmarks Track*. Available at arXiv:2507.01001.

Appendix

A Entrez E-utils Framework

The National Center for Biotechnology Information (NCBI) is a primary repository for genomic data, integrating diverse datasets for genomes, genes, proteins, and more, with analytical and retrieval tools through its website (Ostell, 2012). The Entrez Programming Utilities (E-utilities) (Ostell, 2012), a text-based search system, allows easy navigation across biological databases. Entrez offers a stable interface for querying and retrieving data from 38 NCBI databases, including gene records and biomedical literature (Tatusova, 2010). We make two API calls on the PMC open access subset of the articles: *E-Search*, to find the PubMed IDs (PMIDs) of the most relevant articles for a given query, and *E-Fetch*, to extract the abstracts associated with those PMIDs.

API/Method	Parameters
NCBI API (Entrez)	DB: PubMed, Sort: relevance Min-year: 1990, Max-year: 2024 Max-articles:30, Filter: PMC open access
PaperQA2	Top evidences: 5, Evidence summary length: 50 Max answer source: 3, Answer length: 100 LLM: GPT4o-2024-08-06

Table 7: Parameter settings for NCBI API and PaperQA2

Prompt: Query Reformulation—*SQ* setting

Question: {question}

From the above question, your task is to create a search phrase to find relevant articles. The search phrase should contain only one or a few words. Do not include any word that is not present in the question. Do not generate numbers and punctuation. Convert any initials to their full form.

Below are some examples: {samples}
However, below query term(s) didn't return any article: {query-error}
Your response should only contain the search phrase.
Search phrase:

Table 8: Few-shot LLM prompt to formulate single query terms (*SQ*) to search articles from the NCBI API.

B GeneRAG Settings

GeneRAG retrieves the top 30 articles based on their *relevance* to the query terms and provides them to PaperQA2 (Skarlinski et al., 2024), which uses a ranking technique for creating relevant abstract summaries (to use where LLM input context windows are limited). For MQ_{+Cat}, we generated 3 search terms and concatenated them into a single string for article search. In the case of MQ_{+Stack}, we formulated 10 search terms per question and found 10 abstracts for each term. Then we stacked them in a single list of abstracts to create the context using PaperQA2. In the context produced by PaperQA2, we have top-3 evidence summaries in *SQ* and MQ_{+Cat} settings, and top-5 evidence summaries in MQ_{+Stack} setting. We are using

Prompt: Query Reformulation—*MQ* setting

Your task is to create multiple search phrases to find relevant articles that may assist in answering a question.

1. If the question is simple and asking about one single topic, keep the question as it is.
2. If the question is asking a relation between multiple topics, create sub-questions that will be able to help get the final answer.
3. Create the most important search phrases for the questions to be able to search relevant articles.

The search phrase should contain only one or a few words. Do not include any search term that is not present in the question. Do not generate numbers and punctuations, unless present in the question. Try to convert any initials to their full form.
Below are some samples:

Question: {sample-q}
Reasoning: {sample-sub-questions}
Search phrases: [{sample-query-terms}]
...
However, the query term(s) didn't return any article: {query-error}
Now create n search phrases following the above method for the following question. Your response should contain the sub-question, if created, and the search phrases.

Question:{question}
Your response:

Table 9: Few-shot LLM prompt to formulate multiple query terms (*MQ*) to search articles from the NCBI API.

Prompt: Answer Generation

Question: {question}

For the above question, we retrieve the following context:
Context: {context}

Try to answer the question with the help of the given context in a maximum {length} words with citation in the format (citation key - pmid).

If the context provides insufficient information, try to reply with as much helpful information as you can give. Your response should only contain the answer to the question.
Answer:

Table 10: LLM prompt for generating answers for the given question based on the retrieved context.

Question	Answer (GPT4)
FROSTTOLGENE—What proportion of Transcriptomic studies on plants found ICE1 Transcription factors to be associated with cold acclimation? **[Contains both gene name and trait]**	The provided context indicates that ICE1 transcription factors are associated with cold acclimation in multiple plant studies. Specifically, Wu (2021) discusses the regulation of cold acclimation pathways by ICE1 and ICE2 in plants (Wu 2021 - 34639192). Additionally, Xu (2015) identifies ICE gene homologs in grapevines that play a crucial role in cold acclimation, highlighting the involvement of VaICE1 as a transcription factor responding to cold stress (Xu 2015 - 25019620). However, the exact proportion of studies reporting this association is not specified in the context. Thus, while ICE1 transcription factors are recognized for their role in cold acclimation, quantifying the proportion of studies is not possible based on the given information.
OZWHEATGENE—Candidate gene TraesCS4D03G0067100 on genome GCF_018294505.1 has the annotation of being a GAI-like protein 1 gene. Are there any known associations or links between this family of genes and the trait height in wheat research? Please also describe the evidence for the association. **[Contains both gene name and connection to pre-existing trait]**	Yes, there are known associations between the GAI-like protein family and height traits in wheat. The Reduced height-1 (Rht-B1 and Rht-D1) loci, which are orthologues of the Arabidopsis GAI gene, encode proteins that function as nuclear transcription factors and are implicated in the regulation of plant height by modulating the response to gibberellin, a growth hormone (Peng 1999 - 10421366). The introduction of mutant dwarfing alleles at these loci has led to the development of shorter wheat varieties, significantly contributing to the green revolution's impact on wheat yields (Peng 1999 - 10421366). Additionally, studies of mutant GAI alleles in other crops, such as rice, demonstrate that they also result in reduced gibberellin responses and dwarfism, suggesting that manipulating GAI-like proteins could be a strategy to enhance yield across various species (Peng 1999 - 10421366).

Table 11: Example gene-trait association questions from the FROSTTOLGENE and OZWHEATGENE datasets. Answers are generated by GPT4.

PaperQA2 specifically to find answer-related passages for the answer generation stage. PaperQA2 breaks the retrieved abstracts into smaller chunks and ranks them by their perplexity with respect to the question. Then, the top 5 abstract chunks are selected to generate evidence summaries. Finally, the summaries are further filtered out by re-ranking them by an LLM-prompt. The top 3 of the evidence summaries by this re-ranking were then returned as the final context to be provided to the answer generation stage. The parameter settings for both Entrez system and PaperQA2 are outlined in Table 7.

We also provide the used LLM prompts for query reformulation in *SQ* and *MQ* settings in Table 8 & 9 and answer generation in Table 10. We provide some sample questions and answers for the explored datasets in Table 11.

Query & Context (Prompt Setting–Query LLM)		
Eval. Dim.	FROSTTOLGENE	OZWHEATGENE
$Qr2Q$	MQ_{+Stack}-GPT4	*SQ*-GEMMA
$C2Q$	MQ_{+Stack}-Mistral	MQ_{+Stack}-GPT4
$C2Ab$	*SQ*-GEMMA	MQ_{+Cat}-Mistral
Overall	*SQ*-GPT3.5	SQ-GPT3.5
Answer ((Prompt Setting–Query LLM–Answer LLM)		
$A2Q$	*SQ*-GPT3.5-GPT4	MQ_{+Cat}-Mistral-QWEN
$A2C$	MQ_{+Cat}-Mistral-GPT4	MQ_{+Cat}-GPT4-GPT4
Overlall	MQ_{+Cat}-Mistral-GPT4	MQCAT-GPT4-GPT4

Table 12: Best-performing LLMs for all datasets, based on metric scores computed for multiple evaluation dimensions.

C Metric Details

The evaluation metrics used in this work are as follows:

Semantic similarity measures how closely a text aligns with a reference text in terms of meaning. We chose ROUGE-1 (Lin, 2004) (F1-score), cosine similarity based on SentenceTransformer (Reimers and Gurevych, 2019) embeddings (CosSim), and BERTScore (Zhang et al., 2020), and applied them to $Qr2Q$, $C2Ab$ and $A2C$ evaluation.

Relevance is an LLM-based metric adopted from the RAGAS framework (Es et al., 2024) and measures *Context Relevance* and *Answer Relevance* to the questions for $C2Q$ and $A2Q$ evaluations. *Answer Relevance* is measured by prompting an LLM to create several possible questions that could have led to the given answer. Finally, averagring the cosine similarities between these generated questions and the original question to get the score. *Context Relevance* is measured by asking an LLM to extract sentences from context that are essential for answering the question and computing the ratio of relevant sentences to total context sentences.

Factual consistency evaluates whether the claims in the generated answer can be supported by the source material. We use two metrics: (1) FactCC (Kryscinski et al., 2020), which checks the factual correctness; and (2) AlignScore (Zha et al., 2023) that measures NLI-based textual entailments between two sets of texts. Here, we applied

Prompt	Query	Query to Ques. ($Qr2Q$)					Context to Ques. ($C2Q$)			Context to Abstracts ($C2Ab$)							
Setting	LLM	ROUGE-1	BERTSc.	CosSim.	AlignSc.	Query Avg.	AlignSc.	Relev.	Avg.	ROUGE-1	BERTSc.	AlignSc.	FactCC	Faith.	Avg.	Context Avg.	Overall Avg.
SQ	GPT3.5	0.072	0.877	0.817	0.659	0.606	0.202	0.106	0.154	**0.070**	0.772	0.461	0.879	0.142	0.465	0.376	0.460
	GPT4	0.062	0.877	0.782	0.663	0.596	0.227	0.123	0.175	0.063	0.788	0.487	0.880	0.210	0.486	0.397	0.469
	GEMMA	0.068	0.869	0.824	0.623	0.596	0.217	0.149	0.183	0.065	0.781	0.460	0.883	0.165	0.471	0.389	0.464
	Mistral	**0.089**	0.870	0.758	0.549	0.567	0.229	0.210	0.220	0.047	0.806	0.470	0.839	0.241	0.481	0.406	0.464
MQ_{+Cat}	GPT4	0.029	**0.884**	**0.843**	**0.670**	**0.607**	0.228	0.300	0.264	0.043	0.826	0.455	**0.935**	**0.347**	**0.521**	0.448	**0.505**
	Mistral	0.041	0.870	0.781	0.564	0.564	0.206	**0.329**	0.268	0.048	**0.848**	0.460	0.906	0.324	0.517	0.446	0.489
MQ_{+Stack}	GPT4	0.047	0.855	0.777	0.636	0.579	**0.431**	0.195	**0.313**	0.068	0.811	**0.541**	0.930	0.204	0.511	**0.454**	0.499
	Mistral	0.046	0.855	0.777	0.635	0.578	0.430	0.194	0.312	0.067	0.811	0.540	0.931	0.207	0.511	**0.454**	0.500

Table 13: Metric scores for FROSTTOLGENE query and context evaluation. Queries are formulated by LLMs in three settings, and contexts are produced by PaperQA2. Best scores are **bolded**, and second-best scores are underlined.

Prompt	Query	Answer	Ans. to Ques. ($A2Q$)			Ans. to Context ($A2C$)						
Setting	LLM	LLM	AlignSc.	Relev.	Avg.	AlignSc.	ROUGE-1	BERTSc.	FactCC	Faith.	Avg.	Avg.
SQ	GPT3.5	GPT4	0.710	0.875	0.793	0.676	0.013	0.818	0.345	0.187	0.408	0.518
	GPT4	GPT4	0.719	0.859	0.789	0.695	0.018	0.822	0.254	0.283	0.414	0.521
	GEMMA	GPT4	0.715	0.832	0.774	0.691	0.016	0.821	0.332	0.249	0.422	0.522
	Mistral	GPT4	0.743	0.840	0.792	0.690	0.022	0.831	0.208	0.206	0.391	0.506
	GPT3.5	QWEN	0.758	0.867	0.813	**0.754**	0.012	0.805	**0.736**	0.179	**0.497**	**0.587**
	GPT4	QWEN	0.753	**0.893**	**0.823**	0.730	0.017	0.813	0.684	0.175	0.484	0.581
	GEMMA	QWEN	0.760	0.847	0.804	0.710	0.013	0.814	0.629	0.226	0.478	0.571
	Mistral	QWEN	0.768	0.854	0.811	0.696	0.029	0.824	0.572	0.244	0.473	0.570
MQ_{+Cat}	GPT4	GPT4	0.777	0.798	0.788	0.718	0.042	**0.862**	0.091	0.249	0.392	0.505
	Mistral	GPT4	0.756	0.858	0.807	0.688	**0.043**	0.858	0.265	0.252	0.421	0.531
	GPT4	QWEN	0.716	0.841	0.779	0.659	0.042	0.854	0.060	0.379	0.399	0.507
	Mistral	QWEN	0.717	0.852	0.785	0.653	0.042	0.851	0.311	0.307	0.433	0.533
MQ_{+Stack}	GPT4	GPT4	**0.792**	0.816	0.804	0.746	0.038	0.848	0.088	0.273	0.399	0.514
	Mistral	GPT4	**0.792**	0.831	0.812	0.746	0.038	0.848	0.088	0.273	0.399	0.517
	GPT4	QWEN	0.675	0.821	0.748	0.719	0.038	0.842	0.110	**0.403**	0.423	0.515
	Mistral	QWEN	0.675	0.810	0.743	0.719	0.038	0.842	0.108	0.402	0.422	0.513

Table 14: FROSTTOLGENE metric scores evaluating the answer quality. Answers are generated by GPT4 and QWEN.

the pretrained BERT-based FactCC model[2] and RoBERTa-based *AlignScore-base* model[3]. We applied AlignScore in testing components against the user questions ($Qr2Q$, $C2Q$ and $A2Q$). For $C2Ab$ and $A2C$ evaluation, we applied both FactCC and AlignScore.

Faithfulness refers to the extent to which the LLM's output remains grounded in the evidence. For the $C2Ab$ and $A2C$ evaluations, the information in the context and the answer should be supported by the biomedical evidence given in the reference abstracts and contexts, respectively, without adding irrelevant or hallucinated claims. We used the pretrained hallucination detection model HHEM-2.1-Open model from Vectara[4].

D Automatic Metric Results

We found that there is no single language model that performs well across all stages of the RAG pipeline. As shown in the results in Table 13 on the FROSTTOLGENE dataset, the highest average metric scores were achieved by different models for different evaluation dimensions. For example, for the query-to-question ($Qr2Q$) quality, GPT4 generated queries in MQ_{+Cat} had the best scores. However, GPT4 in MQ_{+Stack} setting generated the best contexts aligned to the question (context-to-question, $C2Q$) and GPT4 in MQ_{+Cat} generated the best contexts with respect to the retrieved abstracts (context-to-abstract, $C2Ab$).

For the answer quality (Table 14), none of the contexts that were generated by the better quality LLMs were able to produce better answers aligned to the question (answer-to-question, $A2C$) or context (answer-to-context, $A2C$). Looking at the summary of the best performing LMs in different settings from Table 12, we observe a similar trend across all datasets. These results essentially indicate that rather than depending on one LLM at each stage of the pipeline, it is better to select a combination of LLMs through multiple evaluation dimensions for optimal outcome.

Furthermore, we present the detailed individual metrics scores and averaged scores on each dimension in Table 13-14 on FROSTTOLGENE and Table 15-16 on OZWHEATGENE datasets. Looking at these detailed scores we observe that, because different evaluation metrics emphasise different

[2] https://huggingface.co/manueldeprada/FactCC
[3] https://huggingface.co/yzha/AlignScore
[4] https://shorturl.at/naxUx

Prompt	Query	Query to Ques. ($Qr2Q$)					Context to Ques. ($C2Q$)			Context to Abstracts ($C2Ab$)							
Setting	LLM	ROUGE-1	BERTSc.	CosSim.	AlignSc.	Query Avg.	AlignSc.	Relev.	Avg.	ROUGE-1	BERTSc.	AlignSc.	FactCC	Faith.	Avg.	Context Avg.	Overall Avg.
SQ	GPT3.5	**0.049**	0.853	0.761	0.544	0.552	0.273	0.028	0.151	0.075	0.753	0.578	0.989	0.054	0.650	0.393	**0.523**
	GPT4	0.046	0.839	0.696	0.483	0.516	0.280	0.085	0.183	0.032	0.760	**0.584**	0.912	0.083	0.634	0.391	0.509
	GEMMA	0.046	0.839	0.696	0.483	0.516	0.280	0.112	0.196	0.032	0.760	**0.584**	**0.995**	0.083	**0.651**	0.407	0.519
	Mistral	0.046	0.851	0.703	0.543	0.536	0.289	0.070	0.180	**0.093**	0.765	0.572	0.898	0.096	0.625	0.398	0.511
MQ_{+Cat}	GPT4	0.036	0.858	**0.806**	0.605	0.576	0.424	0.372	0.398	0.045	0.828	0.494	0.304	**0.317**	0.398	0.398	0.463
	Mistral	0.027	0.852	0.732	0.575	0.547	0.414	**0.402**	0.408	0.084	**0.830**	0.514	0.419	0.226	0.415	**0.413**	0.461
MQ_{+Stack}	GPT4	0.040	**0.861**	0.801	**0.657**	**0.590**	0.488	0.270	0.379	0.043	0.813	0.493	0.205	0.204	0.352	0.359	0.443
	Mistral	0.036	0.840	0.716	0.647	0.560	**0.560**	0.296	**0.428**	0.042	0.810	0.528	0.254	0.163	0.359	0.379	0.445

Table 15: Metric scores for OZWHEATGENE query and context evaluation. Queries are formulated by multiple LLM settings and contexts are produced by PaperQA (GPT4)

Prompt	Query	Answer	Ans. to Ques. ($A2Q$)			Ans. to Context ($A2C$)						
Setting	LLM	LLM	AlignSc.	Relev.	Avg.	AlignSc.	ROUGE-1	BERTSc.	FactCC	Faith.	Avg.	Avg.
SQ	GPT3.5	GPT4	0.678	0.835	**0.757**	0.655	0.006	0.796	0.651	0.080	0.438	0.529
	GPT4	GPT4	0.688	0.798	0.743	0.671	0.006	0.797	0.633	0.091	0.440	0.526
	GEMMA	GPT4	0.688	0.801	0.745	0.671	0.006	0.797	0.633	0.091	0.440	0.527
	Mistral	GPT4	0.666	0.783	0.725	0.667	0.009	0.797	0.416	0.035	0.385	0.482
	GPT3.5	QWEN	0.687	0.816	0.752	0.593	0.004	0.788	0.592	0.089	0.413	0.510
	GPT4	QWEN	0.654	0.818	0.736	0.601	0.005	0.792	**0.656**	0.072	0.425	0.514
	GEMMA	QWEN	0.654	0.820	0.737	0.601	0.005	0.792	0.655	0.072	0.425	0.514
	Mistral	QWEN	0.659	0.813	0.736	0.576	0.008	0.792	0.562	0.052	0.398	0.495
MQ_{+Cat}	GPT4	GPT4	0.620	0.788	0.704	0.765	0.043	**0.874**	0.458	0.396	0.507	0.563
	Mistral	GPT4	0.619	0.796	0.708	0.753	0.040	0.870	0.509	0.396	**0.514**	**0.569**
	GPT4	QWEN	0.649	0.724	0.687	0.651	**0.050**	0.841	0.294	0.464	0.460	0.525
	Mistral	QWEN	0.592	0.722	0.657	0.680	0.036	0.849	0.512	0.483	0.512	0.553
MQ_{+Stack}	GPT4	GPT4	0.687	0.811	0.749	**0.787**	0.039	0.868	0.231	**0.609**	0.507	0.576
	Mistral	GPT4	0.615	0.789	0.702	0.759	0.040	0.863	0.372	0.480	0.503	0.560
	GPT4	QWEN	0.621	0.734	0.678	0.686	0.038	0.845	0.107	0.489	0.433	0.503
	Mistral	QWEN	0.617	0.811	0.714	0.683	0.042	0.845	0.114	0.483	0.433	0.514

Table 16: OZWHEATGENE metric scores evaluating the answer quality. Answers are generated by GPT4 and QWEN.

quality aspects of the evaluated texts, no single metric provides a complete picture. For example, from the FactCC and Faithfulness metric scores for FROSTTOLGENE and OZWHEATGENE datasets in the Table 14 and 16, we observe that the answers generated by using the contexts in the single query setting are factually more accurate, but less faithful than the others. A combination of metrics is therefore more effective, with each metric contributing differently to the assessment of LLM response quality.

Through an analysis of the Pearson correlation coefficient r across different metric combinations, we observe varying levels of correlation between evaluation metrics, as illustrated in the Figure 2. Across all datasets, metrics assessing answer–context alignment (the $A2C$ dimension) exhibit relatively higher correlation with metrics measuring context–question alignment (the $C2Q$ dimension). In contrast, query–question evaluation metrics ($Qr2Q$) show consistently lower correlation with metrics from other evaluation dimensions. Beyond these trends, no consistent correlation patterns are observed across the remaining metric combinations. This analysis indicates that there is no one-size-fits-all metric configuration for evaluating RAG systems, and it is crucial to incorporate human guidance when selecting and combining evaluation metrics.

In Table 17, we report the agreement rates between uniform-weighted metric-based preferences and human preferences (human annotated set) and LLM preferences (synthetic annotations).

E LLM-as-a-Judge

Following the work of Lin et al. (2025), we create in-prompt rubrics containing questions to check the quality of the answers as a guideline for the LLM to generate preferences over pairwise answers. The used rubric questions are listed in Table 18. The generic rubric is created by taking motivation from WildBench (Lin et al., 2025) and LLM-rubric (Hashemi et al., 2024) and connecting with the evaluation dimension explored in our work. To create the domain-specific rubric, we consulted the domain experts.

F Quality Dimensions vs. human reasonings

We prompt *GPT4o-mini* in a one-shot prompt setting to group similar human reasonings and label the quality aspects considered in the reasons. Ta-

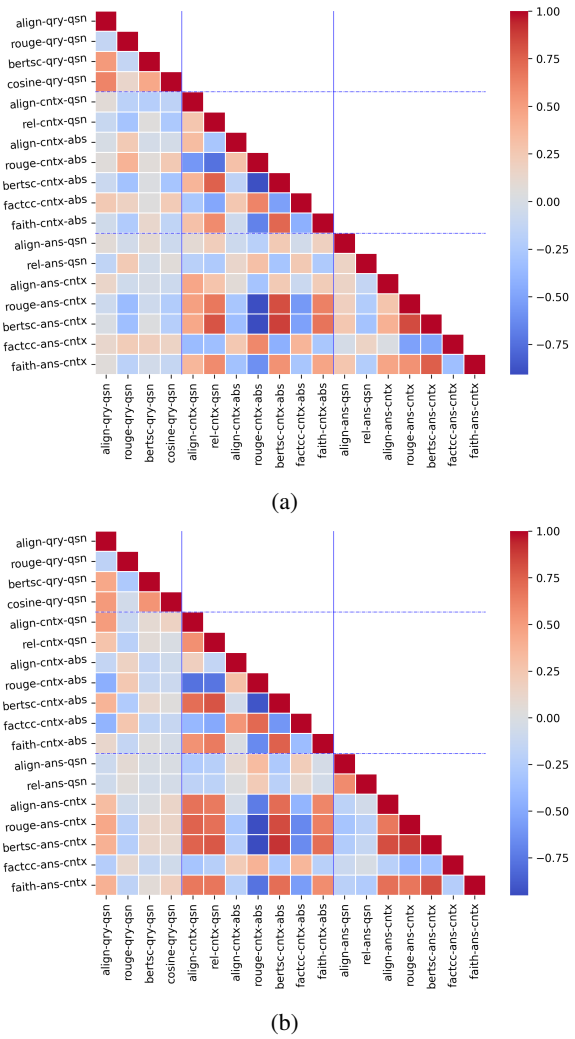

Figure 2: Pearson's correlation coefficients (r) between the evaluation metrics for a) FROSTTOLGENE b) OZWHEATGENE datasets.

Metric	Hum-Metric Agree. (%)	Pearson's r	LLM-Metric Agree. (%)	Pearson's r
All	61.45	**0.300 (p=0.002)**	47.163	**0.267 (p=0.005)**
Qr2Q				
AlignScore	59.04	0.164 (p=0.035)	48.27	-0.044 (p=0.005)
ROUGE	52.41	-0.010 (p=0.902)	46.99	-0.069 (p<0.001)
BERTScore	50.60	-0.003 (p=0.974)	46.77	-0.079 (p<0.001)
CosSim.	51.20	0.002 (p=0.976)	47.85	-0.050 (p=0.001)
C2Q				
AlignScore	56.63	0.094 (p=0.228)	64.37	**0.262 (p<0.001)**
Relevance	51.81	-0.008 (p=0.919)	55.75	**0.082 (p<0.001)**
C2Ab				
AlignScore	53.01	0.032 (p=0.679)	49.23	-0.023 (p=0.138)
ROUGE	49.40	-0.046 (p=0.557)	38.82	-0.226 (p<0.001)
BERTScore	50.60	-0.021 (p=0.790)	54.72	0.055 (p = 0.003)
FactCC	61.45	**0.298 (p=0.010)**	42.77	-0.138 (p<0.001)
Faithfulness	50.60	-0.018 (p=0.822)	57.18	0.106 (p<0.001)
A2Q				
AlignScore	61.64	**0.301 (p=0.050)**	56.42	0.124 (p<0.001)
Relevance	39.73	-0.230 (p=0.051)	43.02	-0.127 (p<0.001)
A2C				
AlignScore	65.75	**0.313 (p=0.007)**	68.00	**0.370 (p<0.001)**
ROUGE	52.05	0.245 (p=0.704)	58.85	0.143 (p<0.001)
BERTScore	57.53	0.253 (p=0.197)	68.20	**0.354 (p<0.001)**
FactCC	57.53	0.245 (p=0.219)	40.83	-0.148 (p<0.001)
Faithfulness	61.64	**0.303 (p=0.039)**	61.82	**0.208 (p<0.001)**

Table 17: Rates of agreements between uniform-weighted metric-based preferences against human preferences (human annotated set) and LLM preferences (synthetic annotations). Bold indicates statistical significance.

Rubric: Generic
- Does the response answer the user's question?
- Is the answer easily understandable and complete with non-repetitive information?
- Are the references useful to answer the question?
- Does the answer cite the references correctly for each claim?
- Can the claims in the answers be accurately supported by the given references?
Rubric: Domain-specific
- Does the answer include a short description of the function of the candidate gene?
- Does the answer state if it can or cannot find evidence linking the candidate gene to the target trait?
- If evidence has been identified, are the type or types of evidence described?
- Does the answer reference the source of all evidence linking the candidate gene to the target trait?

Table 18: Used generic and domain-specific rubrics to guide the LLM in judging the pairwise answers

ble 19 shows the used prompt for this purpose. Some example annotations are illustrated in Tables 23&24. The resultant criteria and the human reasoning under each criterion are presented in Table 22. Based on the criteria labels, the frequency of each criterion in the case of mismatched human-metric criteria for 'all' and a subset of metrics is in Table 21.

G SciArena Dataset

The SCIARENA (Zhao et al., 2025) is an open, crowdsourced benchmarking platform designed to evaluate foundation models on scientific literature-grounded tasks. The dataset[5] contains 13,204 human preferences over pairwise LLM-generated answers on 25 subjects. From this dataset, we investigate metric weight calibration on the subjects closer to our focus domain. The list of subjects and their data size is reported in Table 20 and an example question and answer pair is given in Table 25.

H Human Annotation App

The annotation application was hosted on an internal server. The participants were given a document describing the project, including the annotation task, data collection and anonymisation procedures, and ethics approval, and were asked to voluntarily provide consent via email if they wished to partici-

[5] https://huggingface.co/datasets/yale-nlp/SciArena

Prompt: Grouping Human Ratoinals

human reasonings: {reasons}

Given the above list of reasons for not preferring some LLM-generated answers, try to group them into reasoning categories.

Give a short category name for each category.

Return the output in a JSON firmat, where "criteria" is the "<category label>" and "reasons" is the list of reasons fall into the category.
For example, "criteria": "answer relevance", "reasons": ['The answer is not as relevant as it also does not provide specific information regarding the proportion of studies linking Ethylene-responsive transcription factors with cold acclimation.', 'The answer does not provide any relevant supporting evidence that links ethylene-responsive transcription factors to cold acclimation, making it less relevant.']
Your response:

Table 19: One-shot LLM prompt to group the human reasonings for not preferring an answer and create quality dimension.

Subject	Datasize
Combined	2405
Biomed.	705
Medicine	847
Neuroscience	220
Pharmacy	369
Public Health	264

Table 20: Dataset size of the different subjects in SCIARENA

	Total	Similarity to the Selected Answer	Relevance of Answer	Lack of Details	Lack of Supporting Evidence	Incorrect or irrelevant context	Organization and Structure	Answer Clarity and Fluency	Inaccurate Claims
All	54	16	15	9	4	4	3	2	2
Qr2Q	70	15	28	11	7	4	1	-	1
C2Q	62	18	22	10	6	3	2	-	1
C2Ab	75	17	30	12	7	6	1	1	1
C2Q+C2Ab	70	18	27	11	6	6	-	1	1
A2Q	72	21	26	8	6	5	3	1	2
A2C	56	19	15	8	5	4	2	2	1
A2Q+A2C	57	19	14	9	4	5	3	2	1

Table 21: Criteria from human reasons grouped by LLM for mismatched human-metric preference cases.

pate in the study. We also provided secure logins to the annotators to access the survey application. We presented both answer pairs on a single screen for each question. One answer in the second pair is loaded based on the choice of the first answer. A sample UI showing the portion of the interface is illustrated in Figure 3. The reasons for not preferring an answer are collected as free texts. Also, we employed a separate text box for the annotators to provide stylistic reasons. For ease of reading, we highlighted the same PubMed IDs (PMIDs) with one color.

Criteria	Frequency	Example Reasons
Relevance of Answer	54	'the answer is not as relevant', 'answer is not relevant to the question','the answer is neither sufficiently relevant nor complete.', 'not as relevant to the candidate gene', 'the answer is not as relevant and does not provide inline citations', 'the answer is very similar to the selected answer but less relevant', 'not as relevant does not get to answer', 'not as relevant, no description of gene function', 'the answer is neither sufficiently relevant nor complete.'
Similarity to the Selected Answer	35	'the answer is very similar to the selected answer but not as concise.', 'the answer has less information about the other reference.', 'similar, irrelevant context', 'both bad, no description of gene function', 'the answer is almost similar to the selected answer'
Lack of details	16	'the answer is not as detailed. The paragraph structure is not helpful for directly answer the question.', 'the answer has less information', 'the answer is not as detailed.', 'the answer has less information.', 'details provided could be misleading to a novice.', 'answer is not as detailed', 'less detail', 'the answer is neither sufficiently relevant nor com plete.', 'the answer is not complete'
Incorrect or Irrelevant Context	11	'incorrect context', 'irrelevant context, no description of gene function', 'context has low relevance'
Lack of Supporting Evidence	11	'no supporting evidence', 'the answer does not provide inline citations.', 'the answer is not as relevant and does not provide inline citations', 'did not define function of gene', 'irrelevant; did not define function of gene', 'not as relevant, no description of gene function', 'part of the answer is not as relevant and there is no citations', 'the answer is not as relevant, no supporting evidence'
Organization and Structure	7	'the paragraph structure is not helpful for directly answer the question.', 'the paragraph structure and organization are not as useful', 'the answer is not as relevant paragraph structure and organization are not helpful (not using topic sentence structure)', 'incomplete, intro too long', 'The phrasing in the answer is not as fluent.'
Answer Clarity and Fluency	4	'The phrasing in the answer is not as fluent.', 'Answer did not properly address the two parts of the question.', 'Answer did not fully address the question.', 'The answer is very similar to the selected answer but has less information.The phrasing in the answer is not as fluent.'
Inaccurate Claims	2	'Answer has incorrect claims / evidence', 'not as accurate'
Total reasons	**140**	

Table 22: LLM generated quality dimensions from the human reasons.

Example #1
Question: What proportion of Transcriptomic studies on plants found Gpcr-type g protein 2 gene to be associated with cold acclimation?
Human reason: The answer has less information
Human preference: Answer 2 **Metric preference:** Answer 1
Answer 1: "The provided context does not mention any specific transcriptomic studies on plants that associate the GPCR-type G protein 2 gene with cold acclimation. The cited studies focus on viroid effects on gene expression (Joubert 2022 - 35887025), stress responses via RNA-seq and single-cell RNA-seq (Rurek 2024 - 39524930), and single-cell transcriptomics for plant development (Yu 2023 - 36540021). None of these contexts address GPCR-type G protein 2 gene functionality or its relationship with cold acclimation. Therefore, the proportion cannot be determined.",
Anwer 2: "The provided context does not specify the proportion of transcriptomic studies on plants that found the G-protein coupled receptor (GPCR) type G protein 2 gene to be associated with cold acclimation. It emphasizes the general mechanisms of cold acclimation, such as physiological adjustments and changes in gene expression involving C-repeat binding factors (CBFs) (Barrero-Gil 2019 - 30288701), as well as the role of various proteins in regulating this process (Barrero-Gil 2013 - 23498862). However, specific information regarding the GPCR-type G protein 2 gene's association with cold acclimation is not included in the provided studies. Thus, I cannot provide a definitive answer regarding the proportion."
Example #2
Question: Candidate gene TraesCS6B03G1296700 on genome GCF_018294505.1 has the annotation of being a SKIP interacting protein 16 gene. Are there any known associations or links between this family of genes and the trait height in wheat research? Please also describe the evidence for the association.
Human preference: Answer 1 **Metric preference:** answer 2
Answer 1: "The context does not provide direct evidence linking the SKIP interacting protein 16 gene (TraesCS6B03G1296700) to wheat height. While SKIP is described as a conserved splicing factor and transcriptional regulator in plants (Li 2022 - 35640153), its role in wheat or specific associations with plant height are not mentioned. The TB1 gene in wheat is noted to influence plant height through allelic variation (Dixon 2021 - 32449515), but this is unrelated to SKIP. Additionally, SKIP's role in inhibiting sphingosine kinase (Ghazaly 2020 - 32161116)."
Answer 2: "There is limited direct evidence linking SKIP interacting protein 16 genes, including TraesCS6B03G1296700, to height traits in wheat. However, studies on plant height as an important agronomic trait highlight the role of various genes in influencing this characteristic. The Reduced height-B1b (Rht-B1b) and Rht-D1b alleles have been specifically associated with semi-dwarf plant architecture, enhancing lodging resistance (Song 2023 - 37100915). Furthermore, QTL analyses have been effective in identifying and estimating the effects of genes related to plant height (Paccapelo 2022 - 35597886; Zhou 2023 - 37349717). While the specific role of SKIP interacting protein 16 in height traits remains to be elucidated, its interaction in gene networks may warrant further investigation in the context of height-related QTLs."

Table 23: Example of human annotation and metric-based preference mismatch when two answers are very similar

Question:
Candidate gene TraesCS6B03G0835400 on genome GCF_018294505.1 has the annotation of being a SBP (S-ribonuclease-binding protein) family protein gene. Are there any known associations or links between this family of genes and the trait spike length in wheat research? Please also describe the evidence for the association.
Answer 1:
The SBP (S-ribonuclease-binding protein) family gene TraesCS6B03G0835400 is linked to spike length in wheat. Research indicates that SBP genes, such as BdSBP1 and BdSBP3, play roles in spike development, with their expression patterns during spikelet initiation and responses to environmental factors like vernalization and temperature (Tripathi 2020 - 32929136). A T-DNA mutant of Bdsbp9 showed reduced spike length, highlighting the functional importance of SBP genes in spike development (Tripathi 2020 - 32929136).
Answer 2:
There is currently no direct evidence linking the SBP (S-ribonuclease-binding protein) family genes, such as TraesCS6B03G0835400, to spike length traits in wheat research based on the provided context. The context discusses the association of specific haplotypes in different genes with spike length traits (Zhang 2023 - 37310852) but does not mention SBP family genes or their role in determining spike length. Additionally, while SPL genes are mentioned in relation to plant growth traits (Zhao 2022 - 36011366), they do not directly address the SBP family or spike length in wheat. Therefore, further research would be necessary to explore any potential associations between SBP genes and spike length in wheat.

Annot. 1	Annot. 2
Preference: #1 Rationale: Incorrect claim or evidence	Preference: #1 Rationale: The answer is not as relevant

Table 24: Different reasonings for the same answer pair and the same preferences by two annotators.

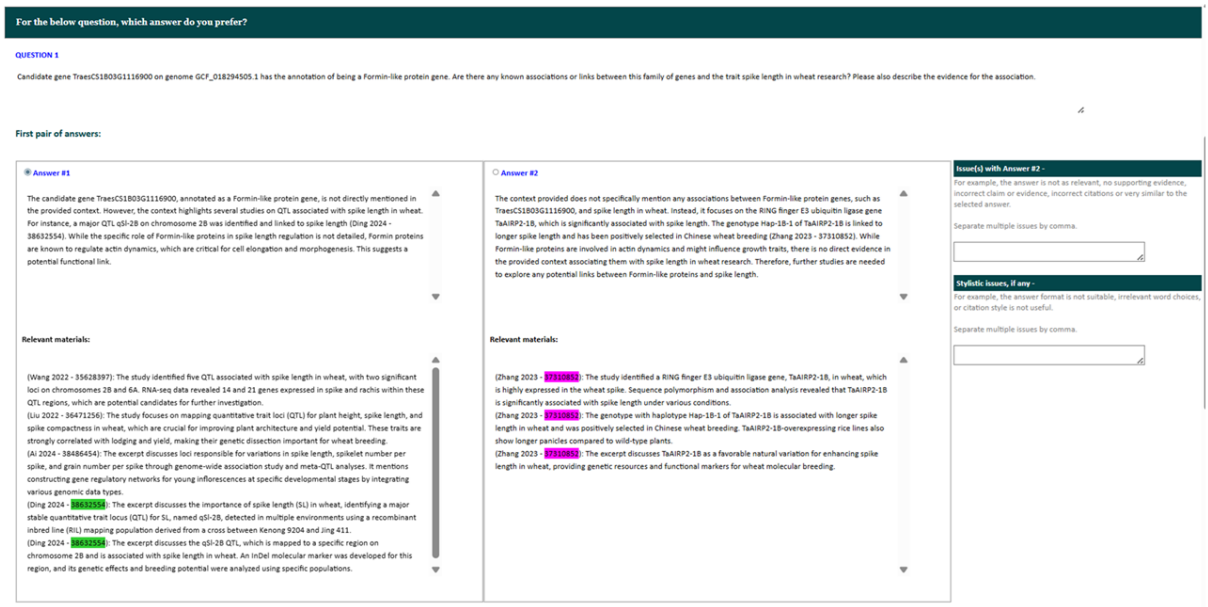

Figure 3: An example UI from the human annotation gathering API.

Subject: Biomed. Engi.

Question: What recent progress characterises novel monolithic stationary phases for liquid chromatography?

Llama-4-Scout: The development of novel monolithic stationary phases has significantly advanced liquid chromatography, offering improved separation efficiency, selectivity, and versatility. Recent progress in this area includes the introduction of nanomaterial-based stationary phases, such as nanoporous silica, zirconia, and graphene, which provide ultra-high surface areas and nanoscale confinement effects (Ratnam et al. 2023b).
Monolithic columns, made from organic polymers, inorganic materials, or hybrid structures, have been engineered to offer tailored characteristics for specific analyses (Nechv0̆0e1talov0̆0e1 et al. 2022a). These columns have demonstrated superior permeability, efficiency, and ease of preparation, making them attractive alternatives to traditional particle-packed columns (Gregu0̆161 et al. 2023a). The incorporation of metal ions, such as Co(II), into molecularly imprinted polymers has enhanced selectivity for separating enantiomeric compounds (Amalia et al. 2022b). Additionally, zwitterionic sulfobetaine-based monoliths have been developed for hydrophilic interaction chromatography, exhibiting good selectivity and efficiency (Wang et al. 2022a).
The use of bio-inspired materials, like lipase-immobilized monoliths, has also shown potential for chiral separations (Fouad et al. 2023a). Furthermore, the integration of monolithic columns with mass spectrometry has enabled sensitive and high-throughput analysis of complex samples (Shamsi et al. 2022a). Overall, these advancements in monolithic stationary phases are transforming liquid chromatography, enabling faster, more sensitive, and high-resolution separations for a wide range of applications (Ratnam et al. 2023b).

Deepseek-R1: Recent advancements in monolithic stationary phases for liquid chromatography (LC) emphasize enhanced selectivity, fabrication techniques, and application-specific designs. Gradient monolithic phases, achieved through controlled chemical modifications like the controlled rate infusion method or photo-grafting, enable gradual changes in surface functionality along the column, improving resolution for complex biological samples (Alla et al. 2022a). Metal-mediated molecularly imprinted polymers (MMIPs) have emerged for chiral separations, with cobalt(II)-based monolithic columns demonstrating high enantioselectivity for racemic citronellal, leveraging metal coordination to enhance template recognition (Amalia et al. 2022b).
Hybrid organic-inorganic monoliths, such as polymer-bound silica phases, combine high permeability with exceptional efficiency (288,000 plates/m for peptides), addressing challenges in biomolecule separation (Ali et al. 2024b). Innovations in miniaturization include microbore columns (1 mm i.d.) that reduce solvent consumption while maintaining performance, ideal for environmentally friendly LC systems (Amalia et al. 2022a). Zwitterionic sulfobetaine-based monoliths, fabricated via rapid photoinitiated copolymerization, offer superior hydrophilic interaction chromatography (HILIC) capabilities, benefiting polar analyte separations (Wang et al. 2022a).
Composite monoliths integrating nanomaterials (e.g., carbon nanotubes) or metal-organic frameworks (MOFs) enhance surface area and selectivity, particularly for proteomic and pharmaceutical analyses (Nechv0̆0e1talov0̆0e1 et al. 2022a). Additionally, advancements in nano-LC utilize silica-based monolithic columns with sub-1 0̆0b5m pores for sensitive food analysis, achieving low back-pressure and high mass transfer kinetics (Aydo0̆11fan et al. 2023a). These developments highlight trends toward multifunctional, application-tailored monoliths, improved fabrication reproducibility, and integration with miniaturized systems, expanding their utility in biomedicine, environmental monitoring, and industrial chromatography (Gregu0̆161 et al. 2023a, Ratnam et al. 2023c).

Human preference: 1

Table 25: Example questions and answers from SciArena dataset for Biomedical Engineering domains