

Human Preference Guided Evaluation of RAG-based QA Systems

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) systems are increasingly used for Question Answering (QA) tasks. However, evaluating their effectiveness remains challenging due to the lack of reliable reference resources and the frequent misalignment between automatic metrics and human judgments. Moreover, there is a scarcity of datasets that contain real-world questions posed by domain experts. To address this problem, we introduce two novel biomedical question datasets concerning the association of wheat genes with specific traits, along with human preference annotations on LLM-generated answers. Methodologically, we explore how to leverage human preferences to calibrate off-the-shelf metrics for automatic evaluation, and we find that the calibrated metrics achieve higher agreement with human preferences compared to baseline metrics on the held-out test set.

1 Introduction

Large Language Models (LLMs) have shown promise in answering biomedical questions (Luo et al., 2022; Bolton et al., 2024; Labrak et al., 2024; Minbyul et al., 2024; Kafkas et al., 2025). Recently, several question answering (QA) models based on Retrieval-Augmented Generation (RAG) techniques have been proposed; for example, Lyu et al. (2024); Jin et al. (2024); Wang et al. (2024) support answer generation, with a special focus on the biomedical domain. However, most of these RAG-based biomedical QA systems are not publicly accessible. Furthermore, they typically focus on answering multiple-choice questions or questions that require short answers (e.g., identifying gene aliases or gene ontology terms). Consequently, the effectiveness of these RAG-based systems in generating long-form responses to biomedical questions remains underexplored.

In a standard RAG setting, given a question, the system first retrieves relevant information as

the *context* and then uses this context to generate an *answer*. However, when evaluating such systems on specialized, discovery-oriented scientific questions, such as “*What proportion of Transcription factors to be associated with cold acclimation?*”, ground-truth answers are usually unavailable. To tackle this challenge, Schimanski et al. (2024) propose evaluating RAG-based QA systems along three quality dimensions: (1) the response should rely only on relevant sources; (2) the response should follow the correct format to maximize traceability; and, (3) the answer should be entailed by the cited source. Similarly, Es et al. (2024) propose a set of evaluation metrics that assess the retrieval system and the answer generation system separately. However, prior works show that experts apply stricter and more domain-specific criteria than non-expert judges (Bavaresco et al., 2025; agu, 2024). It therefore remains unclear whether existing evaluation metrics truly capture the quality dimensions that scientists care about.

To address both of these gaps, we first work with domain experts to build two datasets of plant and agricultural genomic questions (FROSTTOLGENE and OZWHEATGENE), focusing on gene–trait associations in wheat. For these questions, we also collect human preference annotations for model-generated answers, along with the reasons behind their decisions. Then, we examine the correlation between some of the existing automatic evaluation metric scores and the preference judgments provided by domain experts. We find that these metrics, each focusing on different components of the RAG-based system, show varying levels of alignment with human preferences. Inspired by our findings and the recent work on integrating human feedback into finetuning (Bai et al., 2022; Xu et al., 2025) or human-aligned LLM-evaluators (Lee et al., 2024; Ye et al., 2025; Polo et al., 2025), we investigate using human

083 preferences to calibrate existing metrics and find
 084 that the calibrated metrics achieve higher agree-
 085 ment with human preferences than uniform- and
 086 random-weight metrics. Finally, we identify some
 087 of the quality dimensions from human feedback
 088 that are not captured by existing metrics. Our con-
 089 tributions can be summarized as:

- 090 • Two public datasets of plant and agricultural
 091 genomic questions, on gene-trait associations
 092 of the *wheat* that require evidence-supported
 093 answers.¹ We also solicit human preferences
 094 for LLM-generated answers, along with the
 095 reasons behind their decisions.
- 096 • A human preference-guided metric calibration
 097 technique to determine how evaluation met-
 098 rics contribute to the quality assessment of the
 099 LLM-generated responses.
- 100 • Demonstrate the utility of the dataset and the
 101 calibration method in the context of GEN-
 102 ERAG, our proposed system, to answer gene-
 103 phenotype association questions that require
 104 generating *long answers* with evidence from
 105 the literature.

110 2 Problem Statement

111 We assume a 3-stage RAG-based system: (1) query
 112 formulation from user question; (2) retrieval and
 113 corresponding creation of content surrogates (i.e.,
 114 summarizing retrieved documents) for inclusion
 115 in the prompt for answer generation; and, (3) LLM
 116 answer generation. Furthermore, we are assuming
 117 that, in an applied setting, decisions are needed
 118 about which variation is most effective for the pro-
 119 posed use case. Such decisions relate to the choice
 120 of LLM for each component or to algorithmic vari-
 121 ations in any of the three above-mentioned steps.

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

¹Code and datasets: <https://shorturl.at/uTkwm>

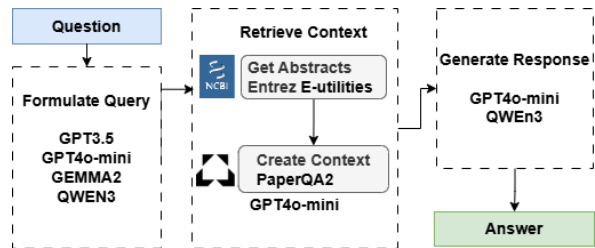


Figure 1: An overview of our RAG-based QA system.

131 3 RAG-based QA System

132 To enable our study, we first built an in-house QA
 133 system, GENERAG, after finding that existing candi-
 134 dates for biomedical QA systems are either not
 135 publicly available or focused solely on generating
 136 short answers. Figure 1 shows the pipeline of our
 137 system and design options for each stage.

138 **Query Formulation.** Our prospective user group,
 139 consisting of plant genomics researchers, typi-
 140 cally relies on NCBI’s Entrez system² to search
 141 databases such as PubMed for relevant scientific
 142 information. To search via the Entrez API (Os-
 143 tell, 2012), we first need to convert a user ques-
 144 tion to search terms. To do this, we prompt LLMs
 145 with few-shot examples to: (1) generate a single
 146 query term for a given question (this setting is de-
 147 noted as SQ); and, (2) create sub-questions from
 148 the given question and then generate a query term
 149 for each sub-question (denoted as MQ). In the SQ
 150 setting, we use the following four LLMs: *GPT3.5-*
 151 *turbo*, *GPT4o-mini*, *GEMMA2-9B*, and *Mistral-*
 152 *Nemo-12B*. In the MQ setting, we use *GPT4o-mini*
 153 and *Mistral-Nemo-12B*.³

154 **Context Retrieval.** GENERAG retrieves rele-
 155 vant abstracts from PubMed focusing only on the
 156 open-access subset. A brief description of the
 157 PubMed API is provided in Appendix §A. In the
 158 MQ setting, where multiple query terms are gen-
 159 erated, two search variants are implemented in GEN-
 160 ERAG: (1) concatenating all query terms into a sin-
 161 gle query string using the OR operator (MQ_{+Cat});
 162 and, (2) retrieving abstracts separately for each
 163 query term (MQ_{+Stack}).

164 After retrieving these abstracts, we follow Pa-

²<https://www.ncbi.nlm.nih.gov/search/>

³We evaluated several domain-specific (BioMistral-7B, BioMedLM, MedLLAMA-7B, PMC-LLAMA-7B and BioGPT-Large) and general-purpose LLMs (QWEN2.5-32B and Claude-3-Sonnet) for query formulation and answer generation on a subset of questions and selected the final models for each stage based on preliminary manual quality inspection.

perQA2 (Skarlinski et al., 2024) pipeline which uses *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 (*GPT4o-mini* or *QWEN-3-8B*) to generate the final answer.

4 Dataset and Human Annotations

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:

- **FROSTTOLGENE**, a set of 42 questions that enquire about the relevance of genes to frost tolerance and cold acclimation, using genes that were identified through a literature search by the domain experts.
- **OZWHEATGENE**, a set of 23 questions from preliminary genome-wide association analysis that identified genes linked to crop yield and component traits for crop yield using 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 questions from the FROSTTOLGENE and OZWHEATGENE datasets and LLM-generated answers through an internally developed survey application (details in Appendix §C).

For each question, we first generated a pool of 12 answers using different variants (i.e., combinations of design options) using our QA system. Next, we selected three candidate answers per question based on the aggregated RAGAS evaluation scores (Es et al., 2024). That is, we sampled the highest-, moderate-, and lowest-quality answers. Finally, these three answers were paired sequentially for annotation. That means, the first pair consisted of two of the three answers chosen randomly, and the second pair consisted of the preferred answer from the first pair and the third answer. In this task, the annotators were not aware of the quality ranks of the

Annotators	Three Answers	Single Pairs
Overall (avg.)	0.50	0.68
E1-E2	0.31	0.53
E2-E3	0.46	0.66
E1-E3	0.74	0.72

Table 1: Inter-annotator agreement (Cohen’s Kappa κ) between the three domain expert annotators (E). Preferences over three answers and single pairs are reported.

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*). We took this approach to reduce annotation effort, while keeping in mind that our focus is on identifying effective evaluation metrics to guide design choices.

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. Ten questions were annotated by three domain expert annotators, and the remaining 53 were single-annotated.

4.2 Inter-annotator Agreements

We measured the inter-annotator agreements under two complementary settings using Cohen’s κ . First, we computed the agreements over the final preferred answer after the annotators (E) had seen all three answers. Second, we computed the agreements over only the identical answer pairs shared across annotators. Table 1 reports the measured agreements. Overall agreement over three-answer preferences was moderate ($\kappa=0.50$), while agreement over identical answer pairs was higher ($\kappa = 0.68$). The agreements varied across annotator pairs, with the highest agreement observed between E1 and E3 ($\kappa=0.739$) for the three-answer preferences, and between E2 and E3 ($\kappa=0.72$) for the common pairs. 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 E1, 0.12 for E2, and 0.32

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

Table 2: List of evaluation metrics selected for evaluation of different evaluation dimensions of GENERAG pipeline. *Qr2Q*=query-to-question, *C2Q*=context-to-question, *A2Q*=answer-to-question, *C2Ab*=context-to-abstract and *A2C*=Answer-to-Context.

for E3, 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. Note, E2 was a bioinformatician, whereas E1 and E3 were plant genomics scientists.

5 Automatic vs. Human Evaluation

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.

To measure the quality of individual components of GENERAG, we selected the semantic similarity, relevance, factual consistency and faithfulness quality dimensions based on our investigation and adopted suitable metrics for each dimension. The metric list is presented in Table 2. Details of the quality dimensions are as follows:

Semantic similarity measures how closely a text aligns with a reference text in terms of meaning. We chose multiple long-standing semantic similarity metrics, such as ROUGE-1 (Lin, 2004) (unigram overlap; F1-score), cosine similarity based on SentenceTransformer (Reimers and Gurevych, 2019) embeddings (CosSim), and BERTScore (Zhang et al., 2020) (contextualised token embeddings), and applied them to *Qr2Q*, *C2Ab* and *A2C* evaluation.

Relevance is a LLM-based metric adopted from the RAGAS framework (Es et al., 2024). We measured the relevance of the retrieved contexts (context relevance) and generated answers (answer relevance) to the user questions for *C2Q* and *A2Q* evaluations. To measure answer relevance, an LLM was prompted to create several possible questions that could have led to the given answer. Finally, the cosine similarities between these generated questions and the original question were computed and averaged to get the answer relevance. Context relevance was evaluated by asking an LLM to extract only the sentences from the context that are essential for answering the question and computing the ratio between the total number of extracted relevant sentences and the total number of sentences in the context.

Factual consistency measures if the generated answer is supported by the source material. In our case, we used two metrics for factual consistency evaluation: (1) FactCC (Kryscinski et al., 2020), which checks the factual correctness; and, (2) AlignScore (Zha et al., 2023) that measures the NLI-based textual entailments between two sets of texts. Here, we applied the pretrained BERT-based FactCC model⁴ and RoBERTa-based *AlignScore-base* model.⁵ Both metrics split the claim and reference texts into small chunks and evaluate each split as a pair. FactCC predicts consistency between pairs, while AlignScore computes a unified alignment score through regression.

We applied AlignScore in testing components against the user questions for *Qr2Q*, *C2Q* and *A2Q* evaluation. 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. In our setting, for the *C2Ab* and *A2C* evaluations,

⁴<https://huggingface.co/manueldeprada/FactCC>

⁵<https://huggingface.co/yzha/AlignScore>

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>
MQ_{+Cat}	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_{+Stack}	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 underlined).

this means that 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. Following RAGAs (Es et al., 2024), we also used the pretrained hallucination detection model HHEM-2.1-Open model from Vectara⁶ as a non-LLM faithfulness measure.

Altogether, we considered 18 evaluation metrics across 5 evaluation dimensions. After evaluating the LLM-generated answers using these metrics, we found that it is not possible to conclusively decide which LLMs and design options to use in the various stages of the RAG system based on the evaluation scores. In other words, none of the variants are consistently performing well across different evaluation dimensions. For example, the results on FROSTTOLGENE in Table 3 showed no consistency on where GPT4 generates the best context. For the $C2Q$ evaluation, the best context was obtained by GPT4 in the MQ_{+Stack} setting, but for the $C2Ab$ evaluation, it was the MQ_{+Cat} setting that obtained the highest results. A more detailed discussion can be found in the Appendix §D.

To contrast our findings with an SOTA dataset, we also generated answers and measured metric scores on the 50 human-gene and disease association questions from GENETURING (Shang et al., 2025), referred to as GTASSOCIATION. We found similar outcomes with this dataset as well (Appendix Tables 16 & 17). This raises our next ques-

⁶<https://shorturl.at/naxUx>

Eval Dim.	Metric	Hum-Metric Agree. (%)	Pearson's r
All		61.45	0.300 (p=0.002)
$Qr2Q$	AlignScore	59.04	0.164 (p=0.035)
	ROUGE	52.41	-0.010 (p=0.902)
	BERTScore	50.60	-0.003 (p=0.974)
	CosSim.	51.20	0.002 (p=0.976)
$C2Q$	AlignScore	56.63	0.094 (p=0.228)
	Relevance	51.81	-0.008 (p=0.919)
$C2Ab$	AlignScore	53.01	0.032 (p=0.679)
	ROUGE	49.40	-0.046 (p=0.557)
	BERTScore	50.60	-0.021 (p=0.790)
	FactCC	61.45	0.298 (p=0.010)
	Faithfulness	50.60	-0.018 (p=0.822)
$A2Q$	AlignScore	61.64	0.301 (p=0.050)
	Relevance	39.73	-0.230 (p=0.051)
$A2C$	AlignScore	65.75	0.313 (p=0.007)
	ROUGE	52.05	0.245 (p=0.704)
	BERTScore	57.53	0.253 (p=0.197)
	FactCC	57.53	0.245 (p=0.219)
	Faithfulness	61.64	0.303 (p=0.039)

Table 4: Rates of alignment between human preference and uniform-weighted metric-based preferences. Bold indicates statistical significance ($\alpha = 0.05$).

tion: if these metrics cannot reach consensus, how can we use them to guide design decisions for the QA system?

5.2 Metric vs. Human Preferences

We start from analyzing the correlation between human preferences and metric-based scores on FROSTTOLGENE and OZWHEATGENE datasets. We converted both human judgments and metric scores into binary preference labels. Formally, for each question q and answer pair (A_a, A_b) , a label of 0 is assigned if annotators prefer A_a , and 1 if they prefer A_b . Metric-based preferences are derived analogously:

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

where $\mathcal{S}(A_i) = \frac{1}{n} \sum_{j=1}^n M_j(A_i)$ calculated over $M_j \in \mathcal{M}$ metrics. Using these labels, we computed the rate of cases when human annotators select the metric-based preferred answer. 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 are reported in Table 4. Overall, experts agreed with the metric-based preferences in 61.45% of the cases (Kappa $\kappa = 0.30$). We also measured a moderate but significant Pearson correlation coefficient, $r = 0.300$, with p-value = 0.002 (significant at $\alpha = 0.01$).

Further, computing the agreement rates between

human preferences and preferences obtained by individual metrics, we found that the agreement levels vary substantially across metrics. When measuring Pearson’s correlations between individual metric-based preferences and human preferences, most metrics do not show statistically significant correlations as well (Table 4). We found *AlignScore* for *A2C* evaluation has the highest human-metric preference agreement of 65.75%. Additionally, *AlignScore* for *A2Q* evaluation achieved 61.64% agreement, *FactCC* for *C2Ab* was 61.45% and *Faithfulness* for *A2C* evaluation achieved 61.64% human-metric preference agreements, with higher statistical significance than the other metrics.

In light of the findings from both metric-based and human evaluations, we infer that not all automatic metrics correlate uniformly with human preferences. In their study [Miller and Tang \(2025\)](#) showed that relying solely on all automated metrics without accounting for their reliability can lead to misaligned evaluations of desired quality criteria and may obscure context-sensitive errors, such as hallucinations and stylistic misalignment. To address this, next, we propose incorporating human preferences into metric reliability calibration by assigning preference-guided metric weights, ensuring human-aligned evaluation of RAG-based QA systems.

6 Preference-guided Metric Weights Calibration

The human–metric preference agreement rates computed using uniformly weighted metrics in §5.2 can be a proxy for human preference alignments. The agreement rates allow us to identify which metrics contribute more strongly to alignment with human judgments. Based on this, we propose a preference-guided approach for calibrating metric reliability weights using human preferences. Our goal is to factor in the relative importance of different metrics aligned with human preferences.

6.1 Proposed Calibration Method

By splitting the annotation dataset into calibration-validation subsets, we considered human-metric preference agreement rates over the calibration set as the calibration weights for the individual metrics. These weights are then used to obtain a weight-adjusted compound metric. Finally, on the validation set, we generated metric-based preferences using this compound metric and compared them

Calib. Technique	Validation Set
Random	55.42
Uniform-weighted	61.44
Calibrated (Ours)	64.30

Table 5: Level of agreements (%) between human and metric-based preferences, computed through random-weights, uniform-weights and calibrated weights on the validation set.

against the human preferences. Our goal was to achieve better alignment with human preferences than the initial human–metric agreement obtained using uniformly weighted metrics on the sampled validation set.

Formally, the agreement rates between human-metric preferences over each metric are considered as the calibration metric weights w_j . Using these weights, the adjusted aggregated scores can be defined as $\mathcal{S}_{adj}(A_i) = \frac{1}{n} \sum_{j=1}^n w_j \times M_j(A_i)$. Then the preference labels $Pref(A_a, A_b | \mathcal{S}_{adj})$ can be obtained using the Eq. 1.

To perform this calibration, we repeated the overall process $k = 100$ times using a 60%–40% random calibration–validation split, resulting in 116 calibration and 50 validation instances per sample. For each k , we sampled the calibration instances and obtained the calibration preferences, \mathcal{C}_{pref} . Then, we computed human–metric preference agreement scores for each metric and applied these agreement scores as metric weights on the validation set. Using the uniformly-weighted metrics on the validation set, we obtain uniform-weighted metric preferences, \mathcal{V}_{unif} and using the weighted-metrics we obtain the weighted-metric preferences, \mathcal{V}_{adj} . The final calibration metric weights for each metric are obtained by averaging the rate of human-metric preference agreements over $k=100$ calibration samples.

For each of the k samples, we also generated random metric weights for the *random-weights* baseline and tested the weights on the validation set by measuring the agreement rates between the human and random-weighted metric preferences.

6.2 Results and Discussion

Our experiments show that calibrating metric weights using human–metric preference agreement leads to a significant improvement in alignment between human and metric-based preferences on the validation set, with an increase in the average agreement rate from 61.44% with uniform metric-weights to 64.30% with calibrated weights.

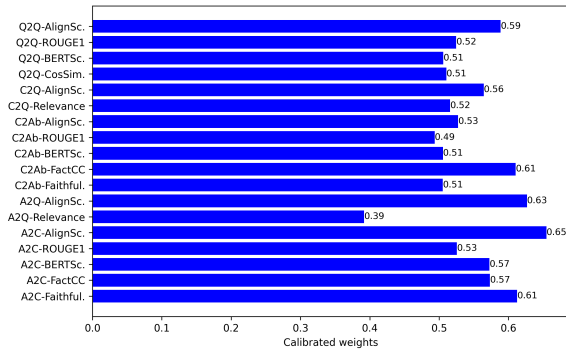


Figure 2: Final calibrated weights for each metric.

This result is also significantly higher than the *random-weights* baseline with an agreement level of 55.42%, as shown in Table 5. The final calibrated metric weights derived by averaging the human-metric preference agreements on the calibration set are shown in Figure 2.

On the validation set, we assess the statistical significance of improvements in alignment with human preferences. For this, we aggregate the \mathcal{V}_{unif} and \mathcal{V}_{adj} preference labels across all k samples and compute Pearson’s correlation with the aggregated human preference labels. We observe a moderate correlation of $r = 0.315$ ($p = 1.6e-47$) between human and \mathcal{V}_{adj} preferences—consistent with the correlation observed in §5.2—compared to a lower correlation of $r = 0.281$ ($p = 9.8e-38$) between human and \mathcal{V}_{unif} preferences.

We also computed agreement rates between human- \mathcal{V}_{unif} and human- \mathcal{V}_{adj} preferences for each k fold. Using Wilcoxon’s signed-rank test (Riffenburgh, 2006), we observe a statistically significant improvement in alignment with human preferences after adjustment ($z = -2.865$, $p = 0.00416$), significant at the $\alpha = 0.005$ level. We further examined per-sample improvements in human alignment by computing the difference in human-metric agreement with and without metric weighting (Figure 3). The orange lines show $Ag_{C-Vunif}$ (human- \mathcal{C}_{pref} vs. human- \mathcal{V}_{unif}), and the blue lines show $Ag_{Vadj-Vunif}$ (human- \mathcal{V}_{adj} vs. human- \mathcal{V}_{unif}). Although in 52% of folds the calibration agreement is lower than the validation agreement (negative $Ag_{C-Vunif}$), applying weights from these folds still improves validation agreement in most cases (positive $Ag_{Vadj-Vunif}$). For $Ag_{Vadj-Vunif}$, we observe that 77% of folds showing improved agreement after weight adjustment (positive $Ag_{Vadj-Vunif}$).

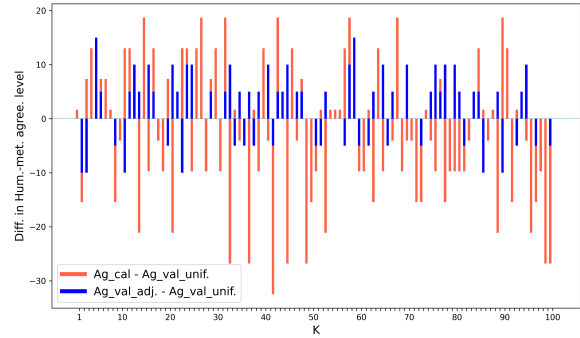


Figure 3: Difference in human-metric agreement across 100 samples ($k = 1 \dots 100$), between the calibration and validation set $Ag_{C-Vunif}$ (orange), and the weight-adjusted and uniform-weighted validation set $Ag_{Vadj-Vunif}$ (blue).

Additionally, to understand which quality dimensions the experts prioritize, we prompted a *GPT4o-mini* model with a one-shot example to group the *human reasonings* and assign interpretable labels (referred to as *criteria*). The prompt is in the Appendix Table 18. From the resultant criteria and associated human reasons, as in the Appendix Table 19, we found that the answer and context relevance (to the question) quality dimensions from our experiment had an intuitive alignment with human annotators’ reasons. In addition to these, human experts also cared about other quality dimensions such as ‘answer completeness’, ‘answer clarity and fluency’ and ‘evidence and citations’. Moreover, they frequently note stylistic issues, such as ‘paragraph structure and organization are not useful’ and ‘does not get to answer’. These findings resonate with prior work that incorporates stylistic and structural qualities (D’Souza et al., 2025), rationale correctness (Jacovi et al., 2024; Bavaresco et al., 2025), and answer usefulness and informativeness (Malaviya et al., 2024) in LLM-based evaluation. Such criteria reflect qualitative aspects of the answers and can be a matter of subjective judgments, varying across expertise levels. Effectively incorporating them, therefore, requires human-in-the-loop evaluation strategies. Using these human-identified criteria as guidance during answer generation may further improve the quality of LLM-generated responses.

7 Related Work

Biomedical QA Datasets. A number of of general and specialized datasets support biomedical QA. MedQA (Jin et al., 2021) and

557 MedRedQA (Nguyen et al., 2023) target general
558 medical knowledge. PubMedQA (Jin et al., 2019)
559 and BIOASQ-QA (Krithara et al., 2023) focus
560 on literature-grounded questions. GeneTuring
561 (Shang et al., 2025) introduces gene-centric bench-
562 mark questions. Multiple-choice datasets such as
563 MMLU (Hendrycks et al., 2021) and LitQA2 (Skar-
564 linski et al., 2024) focus on biomedical and genetics
565 knowledge to generate structured answers. More
566 recently, ExpertQA (Malaviya et al., 2024) curates
567 questions by the domain experts from various fields
568 of study.

569 **LLMs for Question Answering.** LLMs have
570 been pretrained or fine-tuned on biomedical cor-
571 pora to support domain-specific QA, including
572 BioMedLM (Bolton et al., 2024), BioGPT (Luo
573 et al., 2022), PMC-LLAMA (Wu et al., 2024),
574 BioMistral (Labrak et al., 2024), and Bi-
575 oLinkBERT (Yasunaga et al., 2022). Some mod-
576 els further target gene-related tasks, such as Gen-
577 everse (Liu et al., 2024) for gene function un-
578 derstanding and GP-GPT (Lyu et al., 2024) for
579 gene–disease association mapping. In parallel,
580 Retrieval-Augmented Generation (RAG) has been
581 explored for biomedical QA using literature re-
582 trieval (Skarlinski et al., 2024; Garg et al., 2025),
583 structured internal databases (Ning et al., 2025)
584 or combining databases with external tools (Wang
585 et al., 2024). However, such systems rely on re-
586 stricted resources, limiting public availability and
587 reproducibility.

588 **Evaluation of QA systems.** Evaluating QA sys-
589 tems, particularly RAG-based models, remains
590 challenging due to the lack of reliable reference
591 answers. RAGAs (Es et al., 2024) proposes a
592 reference-free evaluation framework judging faith-
593 fulness, relevance, and groundedness. Other works
594 evaluate correctness and faithfulness in instruction-
595 following QA (Adlakha et al., 2024), or factual
596 consistency via metrics such as AlignScore (Zha
597 et al., 2023). Benchmarks such as LitQA2 (Skar-
598 linski et al., 2024) and LabBench (Laurent et al.,
599 2024) assess QA accuracy and biological reasoning
600 skills, while ExpertQA (Malaviya et al., 2024) eval-
601 uates qualitative dimensions including usefulness
602 and cite-worthiness. However, guidance on metric
603 selection for domain-specific QA remains limited.

604 **Human-preferences in LLMs.** Aligning LLMs
605 with human preferences is commonly achieved
606 through Reinforcement Learning from Human

607 Feedback (RLHF) (Bai et al., 2022; Xu et al., 2025).
608 More recently, LLMs have been used as automated
609 evaluators to judge AI-generated responses (Li
610 et al., 2025; Alhawasi and Youssef, 2024; Hijazi
611 et al., 2024; D’Souza et al., 2025). However, stud-
612 ies show that human feedback remains essential
613 for improving evaluator reliability and judgment
614 quality (Miller and Tang, 2025; Bavaresco et al.,
615 2025; Jacovi et al., 2024). Consequently, several
616 approaches now integrate both human and LLM-
617 generated feedback to better align automated eval-
618 uation with human preferences (Yue et al., 2025;
619 Polo et al., 2025; Lee et al., 2024; Ye et al., 2025).

620 8 Conclusions

621 We evaluated the effectiveness of a RAG-based
622 QA system for generating long-form answers to
623 biomedical questions in plant genomics and intro-
624 duced a preference-guided calibration approach
625 for selecting and combining evaluation metrics.
626 To address the lack of non-human genetics QA
627 datasets, we collected two sets of expert-curated
628 plant genomics question datasets and conducted a
629 component-wise evaluation of our RAG pipeline,
630 GENERAG, using diverse reference-free metrics.
631 Our analysis shows substantial variability in how
632 individual metrics align with expert preferences,
633 with no single metric consistently reflecting human
634 judgment.

635 Building on these findings, we proposed a met-
636 ric calibration method that weights evaluation met-
637 rics based on human–metric preference agreement,
638 leading to improved alignment between automated
639 and human evaluations. Analysis of expert ratio-
640 nales further reveals that human judgments also
641 depend on qualitative and stylistic factors beyond
642 commonly used metrics.

643 Future work includes extending this analysis to
644 other biomedical domains, such as clinical deci-
645 sion support and rare disease analysis, to test the
646 generality of preference-guided metric calibration.
647 Another direction is learning preference-aligned
648 metric weights conditioned on question or task
649 types for more adaptive evaluation. Finally, in-
650 tegrating preference-guided metrics into training
651 objectives for RAG-based QA models—through
652 reinforcement learning or self-refinement—could
653 further improve alignment with expert human judg-
654 ments.

9 Limitations

We present a component-wise evaluation of a RAG-based QA system; however, given the many possible design choices at each stage of a RAG pipeline, the evaluation cannot be exhaustive. Due to the complexity of the annotation task, we collected annotations for a limited number of questions, as experts were required to read long answer pairs and supporting abstracts and provide both preferences and justifications. Each annotation took approximately six minutes, and while a larger annotated set would strengthen calibration, this was constrained by annotation effort.

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

2024. [Large language models as financial data annotators: A study on effectiveness and efficiency](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC-COLING*, pages 10124–10145.

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.

Hadel Alhawasi and Abdou Youssef. 2024. [Using llms for evaluating qa systems: Exploration and assessment](#). In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pages 462–469.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

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.

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*.

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 (Volume 1: Long Papers)*, pages 13749–13783.

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.

Mansi Garg, Lee-Chi Wang, Bhavesh Ghanchi, Sanjana Dumpala, Shreyash Kakde, and Yen Chih Chen. 2025. [Biomedical literature q&a system using retrieval-augmented generation \(rag\)](#). *arXiv preprint arXiv:2509.05505*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *9th International Conference on Learning Representations, ICLR*.

Hashem Hijazi, Diego Molla, Vincent Nguyen, and Sarnvaz 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 Trevasakis, 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, Roei 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 (Volume 1: Long Papers)*, pages 4615–4634.

706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760

761	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 . <i>Applied Sciences</i> , 11(14).	818
762		819
763		820
764		
765		
766	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.	821
767		822
768		823
769		824
770		825
771		826
772		
773	Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. 2024. Genegpt: augmenting large language models with domain tools for improved access to biomedical information . <i>Bioinformatics</i> , 40(2).	827
774		828
775		829
776		830
777		831
778	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 . <i>Scientific Reports</i> , 15(1):15093.	832
779		833
780		834
781		835
782		836
783	Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. BioASQ-QA: A manually curated corpus for Biomedical Question Answering . <i>Scientific Data</i> , 10(1):170.	837
784		838
785		839
786		840
787	Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization . In <i>Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)</i> , pages 9332–9346, Online.	841
788		842
789		843
790		844
791		845
792		846
793	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 <i>Findings of the Association for Computational Linguistics: ACL 2024</i> , pages 5848–5864.	847
794		848
795		849
796		850
797		851
798		
799	Jon M Laurent, Joseph D Janizek, Michael Ruzo, Michaela M Hinks, Michael J Hammerling, Sid-dharth Narayanan, Manvitha Ponnampati, Andrew D White, and Samuel G Rodrigues. 2024. Lab-bench: Measuring capabilities of language models for biology research . <i>arXiv preprint arXiv:2407.10362</i> .	852
800		853
801		854
802		855
803		856
804		857
805	Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024. Rlaif vs. rlhf: scaling reinforcement learning from human feedback with ai feedback . In <i>Proceedings of the 41st International Conference on Machine Learning, ICML</i> .	858
806		859
807		860
808		861
809		862
810		863
811		864
812	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 <i>Proceedings of the 31st International Conference on Computational Linguistics</i> , pages 10325–10344.	865
813		866
814		867
815		868
816		869
817		870
		871
		872
		873
	Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries . In <i>Text Summarization Branches Out</i> , 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 <i>Findings of the Association for Computational Linguistics: EMNLP 2024</i> , 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 . <i>Briefings Bioinform.</i> , 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 . <i>arXiv preprint arXiv:2409.09825</i> .	
	Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. ExpertQA: Expert-curated questions and attributed answers . 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 3025–3045.	
	Justin K Miller and Wenjia Tang. 2025. Evaluating llm metrics through real-world capabilities . <i>ArXiv</i> , abs/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 . <i>Bioinformatics</i> , 40:i119–i129.	
	Vincent Nguyen, Sarvnaz Karimi, Maciej Rybinski, and Zhenchang Xing. 2023. MedRedQA for medical consumer question answering: Dataset, tasks, and neural baselines . In <i>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</i> , pages 629–648.	
	Yingpeng Ning, Yuanyuan Sun, Ling Luo, Yanhua Wang, Yuchen Pan, and Hongfei Lin. 2025. Medtrustrag: Evidence verification and trust alignment for biomedical question answering . <i>arXiv preprint arXiv:2510.14400</i> .	
	James M. Ostell. 2012. Entrez: The ncbi search and discovery engine . In <i>Data Integration in the Life Sciences</i> , pages 1–4.	
	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 <i>The Thirty-Ninth Annual Conference on Neural Information Processing Systems</i> .	

874	Nils Reimers and Iryna Gurevych. 2019. SentenceBERT: Sentence embeddings using Siamese BERT-networks . 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 3982–3992, Hong Kong, China.	
875		
876		
877		
878		
879		
880		
881	Robert H. Riffenburgh. 2006. Chapter 6 - statistical testing, risks, and odds in medical decisions . In <i>Statistics in Medicine (Second Edition)</i> , second edition edition, pages 93–114. Academic Press, Burlington.	
882		
883		
884		
885	Tobias Schimanski, Jingwei Ni, Mathias Kraus, Elliott Ash, and Markus Leippold. 2024. Towards faithful and robust LLM specialists for evidence-based question-answering . In <i>Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 1913–1931.	
886		
887		
888		
889		
890		
891		
892	Xinyi Shang, Xu Liao, Zhicheng Ji, and Wenpin Hou. 2025. Benchmarking large language models for genomic knowledge with geneturing . <i>Briefings in Bioinformatics</i> , 26(5).	
893		
894		
895		
896	Michael D Skarlinski, Sam Cox, Jon M Laurent, James D Braza, Michaela Hinks, Michael J Hammerling, Manvitha Ponnampati, Samuel G Rodrigues, and Andrew D White. 2024. Language agents achieve superhuman synthesis of scientific knowledge. <i>arXiv preprint arXiv:2409.13740</i> .	
897		
898		
899		
900		
901		
902	Tatiana Tatusova. 2010. Genomic databases and resources at the national center for biotechnology information . In <i>Data Mining Techniques for the Life Sciences</i> , pages 17–44.	
903		
904		
905		
906	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. <i>arXiv preprint arXiv:2408.01107</i> .	
907		
908		
909		
910		
911	Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama: toward building open-source language models for medicine . <i>Journal of the American Medical Informatics Association</i> , 31(9):1833–1843.	
912		
913		
914		
915		
916	Yifei Xu, Tusher Chakraborty, Emre Kiciman, Bibek Aryal, Srinagesh Sharma, Songwu Lu, and Ranveer Chandra. 2025. RLTHF: targeted human feedback for LLM alignment . In <i>Forty-second International Conference on Machine Learning, ICML</i> .	
917		
918		
919		
920		
921	Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining language models with document links . In <i>Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 8003–8016.	
922		
923		
924		
925		
926	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 <i>The Thirteenth International Conference on Learning Representations, ICLR</i> .	
927		
928		
929		
930		
	Tan Yue, Rui Mao, Xuzhao Shi, Shuo Zhan, Zuhao Yang, and Dongyan Zhao. 2025. QAEval: Mixture of evaluators for question-answering task evaluation . In <i>Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 14717–14730.	931
		932
		933
		934
		935
		936
	Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function . In <i>Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 11328–11348.	937
		938
		939
		940
		941
		942
	Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT . In <i>8th International Conference on Learning Representations, ICLR</i> .	943
		944
		945
		946

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 6: 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 7: 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

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 8: 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 9: LLM prompt for generating answers for the given question based on the retrieved context.

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

972
973
974
975
976
977
978
979
980

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 6 in the appendix.

We also provide the used LLM prompts for query reformulation in *SQ* and *MQ* settings in Table 7 & 8 and answer generation in Table 9. We provide some sample questions and answers for the explored datasets in Table 10.

C 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 participate 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 4. 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.

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 ValCE1 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).
GTASSOCIATION —What are genes related to Chronic atrial and intestinal dysrhythmia? [Contains only disease name]	Genes related to Pseudohypoparathyroidism Ic (PHP-Ic) primarily involve mutations in the GNAS gene, which encodes the alpha-subunit of the stimulatory G protein (Gs α). These mutations lead to diminished Gs α expression or function, resulting in hormonal resistance and the clinical features associated with Albright's hereditary osteodystrophy (AHO). In PHP-Ia/Ic, a heterozygous GNAS1 gene lesion is commonly identified, particularly in a mutational hot-spot involving codons 189-190 (Turan 2016 - 25851935; Linglart 2002 - 11788646). The GNAS complex locus also produces other gene products with mostly monoallelic expression, while Gs α is generally biallelically expressed in most tissues (Turan 2016 - 25851935).

Table 10: Example gene-trait association questions from the three datasets. Answers are generated by GPT4.

For the below question, which answer do you prefer?

QUESTION 1

Candidate gene TraesCS1803G1116900 on genome GCF_018294505.1 has the annotation of being a Formin-like 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.

First pair of answers:

Answer #1

The candidate gene TraesCS1803G1116900, annotated as a Formin-like protein gene, is not directly mentioned in the provided context. However, the context highlights several studies on QTL associated with spike length in wheat. For instance, a major QTL qSL-2B on chromosome 2B was identified and linked to spike length (Ding 2024 - 38632354). While the specific role of Formin-like proteins in spike length regulation is not detailed, Formin proteins are known to regulate actin dynamics, which are critical for cell elongation and morphogenesis. This suggests a potential functional link.

Relevant materials:

(Wang 2022 - 35628397): The study identified five QTL associated with spike length in wheat, with two significant loci on chromosomes 2B and 6A. RNA-seq data revealed 14 and 21 genes expressed in spike and rachis within these QTL regions, which are potential candidates for further investigation.

(Liu 2022 - 36471256): The study focuses on mapping quantitative trait loci (QTL) for plant height, spike length, and spike compactness in wheat, which are crucial for improving plant architecture and yield potential. These traits are strongly correlated with lodging and yield, making their genetic dissection important for wheat breeding.

(Ai 2024 - 38486454): The excerpt discusses loci responsible for variations in spike length, spikelet number per spike, and grain number per spike through genome-wide association study and meta-QTL analyses. It mentions constructing gene regulatory networks for young inflorescences at specific developmental stages by integrating various genomic data types.

(Ding 2024 - 38632354): The excerpt discusses the importance of spike length (SL) in wheat, identifying a major stable quantitative trait locus (QTL) for SL, named qSL-2B, detected in multiple environments using a recombinant inbred line (RIL) mapping population derived from a cross between Kenong 9204 and Jing 431.

(Ding 2024 - 38632354): The excerpt discusses the qSL-2B QTL, which is mapped to a specific region on chromosome 2B and is associated with spike length in wheat. An InDel molecular marker was developed for this region, and its genetic effects and breeding potential were analyzed using specific populations.

Answer #2

The context provided does not specifically mention any associations between Formin-like protein genes, such as TraesCS1803G1116900, and spike length in wheat. Instead, it focuses on the RING finger E3 ubiquitin ligase gene TAAIRP2-1B, which is significantly associated with spike length. The genotype Hap-1B-1 of TAAIRP2-1B is linked to longer spike length and has been positively selected in Chinese wheat breeding (Zhang 2023 - 37310832). While Formin-like proteins are involved in actin dynamics and might influence growth traits, there is no direct evidence in the provided context associating them with spike length in wheat research. Therefore, further studies are needed to explore any potential links between Formin-like proteins and spike length.

Relevant materials:

(Zhang 2023 - 37310832): The study identified a RING finger E3 ubiquitin ligase gene, TAAIRP2-1B, in wheat, which is highly expressed in the wheat spike. Sequence polymorphism and association analysis revealed that TAAIRP2-1B is significantly associated with spike length under various conditions.

(Zhang 2023 - 37310832): The genotype with haplotype Hap-1B-1 of TAAIRP2-1B is associated with longer spike length in wheat and was positively selected in Chinese wheat breeding. TAAIRP2-1B-overexpressing rice lines also show longer panicles compared to wild-type plants.

(Zhang 2023 - 37310832): The excerpt discusses TAAIRP2-1B as a favorable natural variation for enhancing spike length in wheat, providing genetic resources and functional markers for wheat molecular breeding.

Issue(s) with Answer #2 -

For example, the answer is not as relevant, no supporting evidence, incorrect claim or evidence, incorrect citations or very similar to the selected answer.

Separate multiple issues by comma.

Stylistic issues, if any

For example, the answer format is not suitable, irrelevant word choices, or citation style is not useful.

Separate multiple issues by comma.

Figure 4: An example UI from the human annotation gathering API.

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

Table 11: Best-performing LLMs for all datasets, based on metric scores computed for multiple evaluation dimensions.

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 12 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 13), 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 11, 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 12-13 on FROSTTOLGENE, Table 14-15 on OZWHEATGENE and Table 16-17 on GTASSOCIATION datasets. Looking at these detailed scores we observe that due to different evaluation metrics emphasising different 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 13 and 15, 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 5. 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.

E 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. Table 18 shows the used prompt for this purpose. The resultant categories and the human reasonings in the groups are presented in Table 19.

Prompt	Query	Query to Ques. (Qr2Q)					Context to Ques. (C2Q)				Context to Abstracts (C2Ab)					Overall Avg.	
		ROUGE-1	BERTSc.	CosSim.	AlignSc.	Query Avg.	AlignSc.	Relev.	Avg.	ROUGE-1	BERTSc.	AlignSc.	FactCC	Faith.	Avg.		Context Avg.
<i>SQ</i>	GPT3.5	0.072	0.877	0.817	0.659	<u>0.606</u>	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	<u>0.481</u>	0.406	0.464
<i>MQ+Cat</i>	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	<u>0.448</u>	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
<i>MQ+Stack</i>	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	<u>0.312</u>	0.067	0.811	0.540	0.931	0.207	0.511	0.454	<u>0.500</u>

Table 12: 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)						
			AlignSc.	Relev.	Avg.	AlignSc.	ROUGE-1	BERTSc.	FactCC	Faith.	Avg.	Avg.	
<i>SQ</i>	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	
<i>MQ+Cat</i>	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	<u>0.484</u>	<u>0.581</u>	
	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	
<i>MQ+Stack</i>	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	
<i>MQ+Stack</i>	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	<u>0.812</u>	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 13: FROSTTOLGENE metric scores evaluating the answer quality. Answers are generated by GPT4 and QWEN.

Prompt	Query	Query to Ques. (Qr2Q)					Context to Ques. (C2Q)				Context to Abstracts (C2Ab)					Overall Avg.	
		ROUGE-1	BERTSc.	CosSim.	AlignSc.	Query Avg.	AlignSc.	Relev.	Avg.	ROUGE-1	BERTSc.	AlignSc.	FactCC	Faith.	Avg.		Context Avg.
<i>SQ</i>	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	<u>0.407</u>	<u>0.519</u>
	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
<i>MQ+Cat</i>	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	<u>0.408</u>	0.084	0.830	0.514	0.419	0.226	0.415	0.413	0.461
<i>MQ+Stack</i>	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 14: 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)						
			AlignSc.	Relev.	Avg.	AlignSc.	ROUGE-1	BERTSc.	FactCC	Faith.	Avg.	Avg.	
<i>SQ</i>	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	
<i>MQ+Cat</i>	GPT3.5	QWEN	0.687	0.816	<u>0.752</u>	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	
<i>MQ+Stack</i>	GPT4	GPT4	0.620	0.788	0.704	0.765	0.043	0.874	0.458	0.396	0.507	<u>0.563</u>	
	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	<u>0.512</u>	0.553	
<i>MQ+Stack</i>	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 15: OZWHEATGENE metric scores evaluating the answer quality. Answers are generated by GPT4 and QWEN.

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.
<i>SQ</i>	GPT3.5	0.052	0.895	0.861	0.662	<u>0.618</u>	0.362	0.445	<u>0.404</u>	0.045	0.806	0.465	0.919	0.114	0.470	0.451	0.511
	GPT4	0.049	0.898	0.877	0.667	<u>0.623</u>	0.332	0.455	0.394	0.039	0.808	0.454	0.925	0.132	0.472	0.450	0.512
	GEMMA	0.056	0.900	0.892	0.683	0.633	0.343	0.461	0.402	0.040	0.802	0.462	0.924	0.110	0.468	0.449	0.516
	Mistral	0.053	0.881	0.750	0.518	0.551	0.271	0.390	0.331	0.045	0.841	0.538	0.914	0.396	<u>0.547</u>	<u>0.485</u>	0.509
<i>MQ+Cat</i>	GPT4	0.056	0.878	0.862	0.637	<u>0.608</u>	0.324	0.474	0.399	0.061	0.830	0.506	0.933	0.368	0.540	0.449	0.539
	Mistral	0.058	0.876	0.849	0.572	<u>0.589</u>	0.299	0.362	<u>0.331</u>	0.062	0.835	0.511	0.924	0.410	0.548	0.486	<u>0.523</u>
<i>MQ+Stack</i>	GPT4	0.054	0.851	0.762	0.531	<u>0.550</u>	0.450	0.379	0.415	0.042	0.814	0.536	0.289	0.177	0.372	0.384	0.444
	Mistral	0.057	0.858	0.789	0.554	<u>0.565</u>	0.437	0.327	<u>0.382</u>	0.047	0.811	0.526	0.311	0.189	0.377	0.378	0.446

Table 16: Metric scores for GTASSOCIATION 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.	Answer Avg.	
<i>SQ</i>	GPT3.5	GPT4	0.570	0.765	0.668	0.778	0.039	0.803	0.755	0.879	<u>0.650</u>	0.656	
	GPT4	GPT4	0.572	0.777	0.675	0.766	0.037	0.889	0.536	0.882	0.622	0.637	
	GEMMA	GPT4	0.545	0.763	0.654	0.749	0.039	0.805	0.677	0.815	0.617	0.628	
	Mistral	GPT4	0.563	0.759	0.661	0.673	0.042	0.882	0.598	0.577	0.554	0.585	
	GPT3.5	QWEN	0.588	0.763	0.676	0.674	0.039	0.891	0.602	0.681	0.577	0.605	
	GPT4	QWEN	0.597	0.789	<u>0.693</u>	0.641	0.041	0.890	0.515	0.737	0.565	0.601	
<i>MQ+Cat</i>	GPT4	QWEN	0.582	0.797	0.690	0.666	0.038	0.894	0.577	0.684	0.572	0.605	
	Mistral	QWEN	0.601	0.747	0.674	0.605	0.040	0.869	0.537	0.513	0.513	0.559	
	GPT4	GPT4	0.531	0.786	0.659	0.747	0.037	0.902	0.816	0.813	0.663	0.662	
	Mistral	GPT4	0.540	0.775	0.658	0.720	0.037	0.894	0.589	0.769	0.602	0.618	
	GPT4	QWEN	0.596	0.780	0.688	0.657	0.040	0.889	0.663	0.657	0.581	0.612	
	Mistral	QWEN	0.589	0.808	0.699	0.652	0.040	0.883	0.577	0.685	0.567	0.605	
<i>MQ+Stack</i>	GPT4	GPT4	0.523	0.781	0.652	0.710	0.039	0.881	0.511	0.851	0.598	0.614	
	Mistral	GPT4	0.537	0.793	0.665	0.666	0.037	0.874	0.567	0.783	0.585	0.608	
	GPT4	QWEN	0.587	0.788	0.688	0.629	0.038	0.866	0.425	0.654	0.522	0.570	
	Mistral	QWEN	0.582	0.792	0.687	0.628	0.036	0.861	0.468	0.665	0.532	0.576	

Table 17: GTASSOCIATION metric scores evaluating the answer quality. Answers are generated by GPT4 and QWEN.

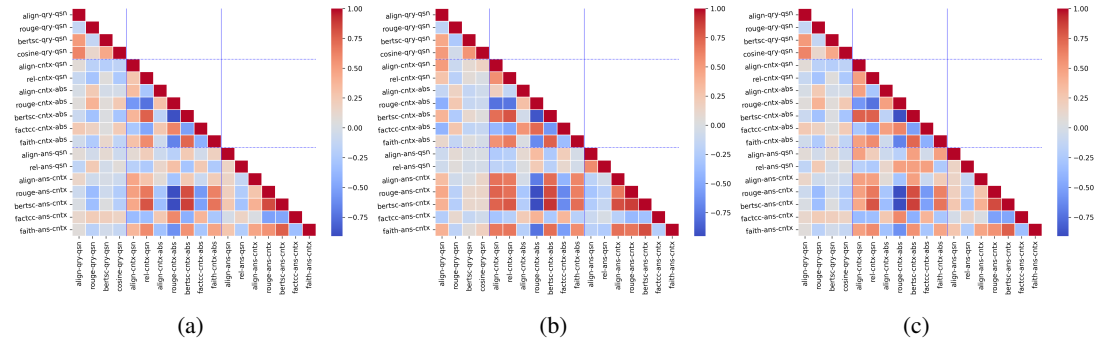


Figure 5: Pearson’s correlation coefficients (r) between the evaluation metrics for a) FROSTTOLGENE b) OZWHEAT-GENE and c) GTASSOCIATION datasets.

Prompt: Grouping of the human reasonings

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 format, 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 18: One-shot LLM prompt to group the human reasonings for not preferring an answer and create quality dimension.

Criteria	human reasonings
Answer relevance	‘The answer is neither sufficiently relevant nor complete.’, ‘The answer is not as relevant; paragraph structure and organization are not as useful.’, ‘the answer is not as relevant; no supporting evidence.’, ‘the answer is not as relevant; the logic of the paragraph does not sound.’, ‘the answer is not relevant to the question.’, ‘The answer is not as relevant and does not provide inline citations.’, ‘the answer is not relevant to the question.’, ‘The answer does not properly address the question raised.’, ‘Answer #2 deviated from the question asked.’, ‘Answer #1 provided answers not related to the question asked.’, ‘Answer is irrelevant to the question.’, ‘Details provided could be misleading to a novice.’, ‘The answer is not relevant to the question.’, ‘The answer is not relevant to the question.’, ‘Answer #2 listed studies/research methods irrelevant to the question.’
Answer completeness	‘The answer is not complete; paragraph structure and organization are not useful.’, ‘the answer is very similar to the selected answer but has less information.’, ‘the answer is very similar to the selected answer but not as concise.’, ‘the answer is not as detailed.’, ‘the answer has less information.’, ‘the answer is not as relevant; the paragraph structure is not helpful.’, ‘the answer is not as relevant, no description of gene function.’, ‘the answer is not as relevant; does not get to answer.’, ‘not as relevant; does not get to answer, no description of gene function.’, ‘not as relevant; did not define function of gene.’
Answer clarity and fluency	‘The phrasing in the answer #1 is not as fluent.’, ‘The paragraph structure is not helpful for directly answering the question.’, ‘Answer #2 is typically not so user-friendly with the repetitive phrasing of the question.’, ‘Answer #2 did not properly address the two parts of the question.’, ‘Answer #1 did not fully address the question.’
Evidence and citations	‘no supporting evidence.’, ‘The answer does not provide inline citations.’, ‘contains irrelevant citations.’, ‘incorrect claim or evidence, not as relevant.’
Similarity to other answers	‘the answer is very similar to the selected answer.’, ‘the answer is the same to the selected answer.’, ‘Answer #1 is very similar to answer #2.’, ‘Answer #2 is very similar to the selected answer.’, ‘both answers are very similar, and both tackled the question well.’
Context relevance	‘context has low relevance.’, ‘The last paragraph of Answer #2 veered off the question.’, ‘although suggesting other genes linked to frost tolerance could be useful in some cases, it is best to stick to the question asked.’

Table 19: LLM generated quality dimensions from the human reasonings.