Beyond BLEU: Ethical Risks of Misleading Evaluation in Domain-Specific QA with LLMs

Ayoub Nainia[1] **Régine Vignes-Lebbe**[1] **Hajar Mousannif**[2] **Jihad Zahir**[2,3]
[1]Institut de Systématique, Evolution, Biodiversité (ISYEB), Sorbonne Université, Muséum National d'Histoire Naturelle, CNRS, EPHE-PSL, Université des Antilles, F-75005, Paris, France
[2]LISI Laboratory, Cadi Ayyad University, Marrakesh, Morocco
[3]UMMISCO, IRD, France
ayoub.nainia@sorbonne-universite.fr,
regine.vignes_lebbe@sorbonne-universite.fr,
mousannif@uca.ac.ma, j.zahir@uca.ac.ma

Abstract

Large Language Models (LLMs) are increasingly used in scientific question answering (QA), including high-stakes fields such as biodiversity informatics. However, standard evaluation metrics such as BLEU, ROUGE, Exact Match (EM), and BERTScore remain poorly aligned with the factual and domain-specific requirements of these tasks. In this work, we investigate the gap between automatic metrics and expert judgment in botanical QA by comparing metric scores with human ratings across five dimensions: accuracy, completeness, relevance, fluency, and terminology usage. Our results show that standard metrics often misrepresent response quality, particularly in the presence of paraphrasing, omission, or domain-specific language. Through both quantitative analysis and qualitative examples, we show that high-scoring responses may still exhibit critical factual errors or omissions. These findings highlight the need for domain-aware evaluation frameworks that incorporate expert feedback and raise important ethical concerns about the deployment of LLMs in scientific contexts.

1 Introduction

Large language models (LLMs) are increasingly fine-tuned and deployed for question answering (QA) in specialized domains such as biodiversity, medicine, and scientific research. These models offer compelling fluency and broad generalization capabilities, making them attractive for automating knowledge access in fields where information is complex and rich in terminology. However, evaluating their effectiveness in the real-world, especially in high-stakes contexts, remains a critical challenge.

Despite impressive reported performance, most QA systems are evaluated using lexical overlap metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), or Exact Match (EM) (Rajpurkar et al., 2016a). These metrics, while easy to compute, have well-documented limitations: they reward surface similarity over factual accuracy, fail to penalize hallucinated content, and systematically favor longer redundant answers that may appear plausible but lack precision (An et al. (2024b); Maynez et al. (2020)). In scientific and technical domains where answers must be both correct and complete, such metrics can inflate perceived performance and mask serious factual deficiencies.

This issue is especially pronounced in domain-specific Question Answering (QA), where small inaccuracies, such as an incorrect botanical trait or a misrepresented medical guideline, can undermine the reliability of the entire system. Recent studies in medical QA (Singhal et al. (2023); Moor et al. (2023)) and scientific QA (Taylor et al., 2022a) demonstrate that even fine-tuned LLMs often generate answers that sound correct but are either partially wrong, incomplete, or not grounded in verifiable sources. However, these limitations are rarely visible in standard evaluation scores, leading to misguided claims about model readiness and potential misuse in real-world deployments.

In this paper, we critically examine how current evaluation practices contribute to an overestimation of fine-tuned LLM performance in domain-specific QA tasks. Our analysis focuses on botanical trait extraction, a high-stakes scientific application where factual precision and accurate use of terminology are essential. However, the evaluation challenges we highlight are not limited to botany. They also apply to fields such as medicine and law, where even small factual errors can have serious consequences (Singhal et al., 2022; Weidinger et al., 2021). In legal contexts, for example, recent efforts have emphasized the importance of expert-annotated datasets and domain-tuned models to ensure accurate interpretation of statutes and

Proceedings of the First Workshop on Comparative Performance Evaluation: From Rules to Language Models associated with RANLP 2025, pages 77–86, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-102-8-009

regulations (Al Mouatamid et al., 2023).

Biodiversity data, for example, serves as the foundation for ecological research, conservation policy, endangered species monitoring, and climate impact studies. Errors in trait extraction can propagate into global biodiversity databases such as the Global Biodiversity Information Facility (GBIF)[1], leading to misclassifications, flawed scientific conclusions or misinformed policy decisions. Even minor hallucinations or omissions (e.g., in leaf morphology or species distribution) can distort downstream analysis or fieldwork.

We analyze cases where model outputs receive high automatic scores but fail expert evaluation due to factual inaccuracies, incompleteness, or loss of critical context. We propose a set of evaluation principles for scientific QA that prioritize factual faithfulness, information coverage, and grounding in verifiable sources, dimensions often invisible to surface-level metrics like BLEU or EM.

Our findings highlight the need to move beyond BLEU and toward evaluation frameworks that reflect the true utility and limits of LLMs in high-precision domains.

2 Background and Motivation

Automated question answering (QA) systems, including fine-tuned large language models (LLMs), are commonly evaluated using lexical overlap metrics originally developed for tasks such as machine translation and summarization. Among the most widely adopted are BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and Exact Match (EM) (Rajpurkar et al., 2016a) and token-level F1 from extractive QA benchmarks such as SQuAD (Rajpurkar et al., 2016a). These metrics compute n-gram overlap between system outputs and reference answers, either as precision (BLEU), recall (ROUGE), or strict equality (EM). Their popularity is largely due to their ease of implementation, reproducibility, and long-standing use in benchmark comparisons.

However, a growing body of work has questioned the adequacy of these metrics in generative QA settings, where answers are open-ended, multi-sentence, and potentially phrased in ways not captured by reference strings. BLEU and ROUGE focus on surface-level n-gram similarity and do not assess whether an answer is factually correct, complete, or grounded. For example, Maynez et al. (2020) showed that summarization models frequently hallucinate content that is not supported by the source text, but still receive high ROUGE scores. An et al. (2024b) found similar trends in long-form QA: answers that are fluent but incorrect or incomplete are often rewarded by BLEU and ROUGE, while semantically valid but lexically diverse answers are penalized. These findings are echoed in previous critiques (Yang et al., 2018) that demonstrated that metrics such as BLEU and ROUGE poorly capture answer quality in both yes/no and entity-centric QA formats.

Despite these limitations, overlap-based metrics remain dominant, including in domain-specific QA systems. Biomedical, legal, and scientific QA models routinely report BLEU, ROUGE, and Exact Match (EM) as primary evaluation metrics (Lee et al., 2021) (Singhal et al., 2023), often without rigorous human evaluation or claim-level verification (Thorne et al., 2018). In practice, this can lead to inflated perceptions of model performance, especially when answers contain hallucinated or missing information that metrics fail to penalize. This risk is amplified in high-stakes domains such as medicine or biodiversity science, where users may trust a model's fluent output without realizing that it lacks factual correctness or critical details.

The continued reliance on these metrics presents not only a technical concern, but an ethical one Ferdaus et al. (2024). By overstating model reliability, current evaluation practices may contribute to misleading claims of safety and readiness, potentially enabling misuse or over-deployment in sensitive contexts. As LLMs are increasingly proposed as tools for scientific assistance and clinical support, evaluating them using metrics that do not reflect truthfulness, completeness, or verifiability is insufficient and potentially dangerous.

3 Related Work

As large language models (LLMs) are increasingly deployed in high-stakes domains, their evaluation has become a focal point of methodological concern and ethical debate. This section reviews work on QA evaluation metrics, factuality assessment, domain-specific QA challenges, and the responsible deployment of LLMs. Our contribution builds on these foundations by examining how inadequate metrics can systematically misrepresent the real capabilities of fine-tuned models in scientific contexts.

[1]https://www.gbif.org

3.1 Evaluation Metrics for LLM Question Answering

Traditional QA evaluation is heavily based on n-gram overlap metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and Exact Match (EM) (Rajpurkar et al., 2016b). However, these metrics often fail to capture semantic correctness, factual consistency, or completeness, especially in open-ended or domain-specific QA tasks. Studies have shown that token-level overlap correlates poorly with human judgment on complex QA (Yang et al., 2018; Maynez et al., 2020)). The L-Eval benchmark introduced by An et al. (2024a) further demonstrated that BLEU and ROUGE do not align with human preferences, particularly in long-context reasoning tasks. They are also biased toward verbose, lexically similar outputs, further inflating scores for answers that may be inaccurate or incomplete.

More sophisticated semantic similarity metrics, such as BERTScore (Zhang et al., 2020), address paraphrastic variation, but remain sensitive to domain-specific terminology and formatting. In response, some QA evaluations now combine overlap-based metrics with embedding-based similarity and human assessment. There is also increasing interest in using LLMs themselves as evaluators (LLM-as-a-judge) (Zheng et al., 2023), although these introduce new biases (Dubois et al., 2025). In general, there is a growing consensus that surface-level metrics are insufficient to capture factual accuracy in generative QA.

3.2 Factuality and Hallucination Detection in LLMs

In light of these limitations, recent research has focused on factuality: whether generated answers are supported by verifiable evidence. Maynez et al. (2020) and Pagnoni et al. (2021) found that summarization systems often hallucinate content while scoring highly on ROUGE. These findings motivated the development of claim-level verification benchmarks such as FRANK (Pagnoni et al., 2021), HaDeS (HAllucination DEtection dataSet) (Liu et al., 2022), and TruthfulQA (Lin et al., 2022), which assess hallucination at the token level.

Several methods now use retrieval-augmented QA (Lewis et al., 2021), natural language inference, or question decomposition to verify generated content (e.g., QAFactEval (Fabbri et al., 2022), RefChecker (Hu et al., 2024), Attributable to Identified Sources (AIS) (Rashkin et al., 2023)). However, even retrieval-based systems can hallucinate when the retrieved content is incomplete or ambiguous (Moor et al., 2023). Hence, detection remains challenging, with token-level hallucination detectors achieving only $\sim$70% F1 in specialized domains, indicating that hallucination remains a persistent issue even with dedicated detectors.

3.3 Evaluation in Domain-Specific QA Systems

Evaluating QA systems, in scientific domains, introduces unique challenges. Domain-specific LLMs such as Microsoft's BioGPT (Luo et al., 2022) and Meta's Galactica (Taylor et al., 2022b) perform well on tailored benchmarks but require expert-informed evaluation to ensure factual grounding (Singhal et al., 2023; Bélisle-Pipon, 2024). In medicine, for example, Med-PaLM's evaluation combined human review with metrics to assess not just correctness, but also reasoning quality, potential harm, and trustworthiness (Singhal et al., 2022). However, hallucinations and omissions persisted, where LLMs struggled to contextualize general knowledge into actionable recommendations.

In botany and biodiversity informatics, research is emerging on LLM-based extraction of scientific facts from unstructured text, with recent studies achieving over 90% precision in tasks such as species identification, geocoding, and data structure (Castro et al., 2024). However, these results often mask persistent challenges that standard metrics fail to capture. Current LLMs show a concerning tendency for delivering incorrect information that raises concerns about their reliability in ecological research applications (Gougherty and Clipp, 2024). At the same time, extracting structured knowledge from scientific text remains fundamentally challenging even for fine-tuned models (Dagdelen et al., 2024).

The field faces several domain-specific obstacles that standard metrics do not address. Term ambiguity represents a major challenge, as ecological and botanical terminology often carries context-dependent meanings that LLMs struggle to disambiguate correctly. Domain-specific syntax further complicates extraction, as scientific literature employs specialized linguistic patterns and taxonomic conventions that differ markedly from general text. Additionally, the propagation of subtle errors poses particular risks in scientific contexts, where small inaccuracies can compound into significant misrepresentations of ecological relationships or species characteristics.

Perhaps most critically, current benchmarks often fall short in capturing the diverse behavior of these models in real-world applications, with existing frameworks being limited by their focus on general-purpose queries and lack of diversity across specialized domains (Raju et al., 2024). The absence of curated benchmarks specifically designed for biodiversity informatics, combined with limited human-in-the-loop evaluation frameworks, makes it difficult to reliably assess model factuality, completeness, or risk of systematic errors in scientific knowledge extraction. Although domain-specific datasets such as FloraNER have emerged for botanical named entity recognition (Nainia et al., 2024), these represent only narrow aspects of the broader challenge of biodiversity informatics, leaving significant evaluation gaps in other critical areas such as ecological relationship extraction, species behavior analysis, and cross-domain knowledge integration. This evaluation gap is particularly concerning given that scientific problem-solving requires domain expertise, understanding of long-context information, and multi-step reasoning (Cui et al., 2025) that may not be adequately tested by existing metrics (Dorm et al., 2025).

3.4 Responsible Use and Deployment in High-Stakes Domains

Ethical concerns about LLM deployment have intensified in law, science, and medicine, where overreliance on fluent but inaccurate outputs has led to misinformation, bogus citations, and incorrect legal filings (Weidinger et al., 2021). Therefore, Scholars have called for stricter evaluation, transparency, and oversight, especially for systems supporting scientific reasoning or clinical advice (Bélisle-Pipon 2024; Giorgino et al. 2023).

Safeguards such as retrieval-augmented generation (RAG) (Chen et al., 2024), expert-led evaluation, and alignment methods like reinforcement learning from human feedback (RLHF) (Christiano et al., 2023) exist, but are inconsistently applied. Many published evaluations still rely heavily on lexical overlap metrics.

While prior work notes the risks of hallucination and the limits of automatic metrics, few studies have examined these failures in domain-specific QA. Our work addresses this gap through targeted failure analyses and by proposing ethically grounded evaluation principles centered on factuality, completeness, and verifiable grounding, better reflecting the needs of high-stakes scientific tasks.

4 Methodology

To investigate the ethical limitations of automatic evaluation metrics in domain-specific question answering (QA), we analyze a French-language QA system fine-tuned on botanical texts. The system follows a two-stage architecture: the first stage generates trait-specific questions from floristic descriptions, and the second stage answers those questions based on the same context.

In our experiments, we evaluate only the answer generation stage, as factual reliability and completeness are most critical for downstream use. Question generation is included to relieve users from formulating queries and to provide standardized prompts, but its intrinsic quality was not separately assessed.

The QA system is based on LLaMA 2[2] and LLaMA 3[3] models, fine-tuned using Low-Rank Adaptation (LoRA) (Hu et al., 2021). The training dataset consists of 16,962 expert-verified question–answer pairs constructed from unstructured botanical descriptions. Each QA pair is associated with a specific botanical trait (e.g., leaf shape, flower color, inflorescence length) and was designed to reflect structured knowledge retrieval from naturalistic text.

4.1 Evaluation Dataset

For evaluation, we curated a held-out test set of 1,697 botanical contexts from a distinct source corpus not used during training. From this, a representative sample of 100 model outputs was randomly selected using a reproducible pandas-based function for expert-based review. Each sample consists of a botanical description (context), a trait-specific question, and the system-generated answer

4.2 Human Evaluation Protocol

Each of the 100 outputs was independently reviewed by a biodiversity expert. The expert rated each answer on a 1-5 Likert scale across five dimensions: Accuracy, Completeness, Relevance, Fluency, and Terminology Usage (Table 1).

[2]https://huggingface.co/meta-llama/Llama-2-7b
[3]https://huggingface.co/meta-llama/Meta-Llama-3-8B

Metric	Meaning
Accuracy	Factual correctness of the response
Completeness	Inclusion of all relevant information from the context
Relevance	Appropriateness of the answer given the question
Fluency	Grammatical and stylistic quality
Terminology Usage	Correct and domain-appropriate terminology

Table 1: Expert evaluation metrics for assessing response quality in domain-specific QA.

The expert was provided with the meaning of each evaluation expert-based metric to ensure consistent scoring in all examples.

4.3 Automatic Metrics

To assess how standard metrics reflect answer quality, we computed the following scores for the same 100 examples: BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), BERTScore (Zhang et al., 2020), and Exact Match (EM) (Rajpurkar et al., 2016a). These metrics were computed using the model outputs and their corresponding reference responses from the training set. We then compared these automatic scores to human ratings in order to analyze discrepancies and identify failure cases that raise ethical concerns.

4.4 Ethical Framing

This methodology is designed to reveal how surface-level metrics such as BLEU and ROUGE may produce inflated scores for outputs that are fluent, but factually incorrect, incomplete, or misleading. In scientific domains such as botany, such evaluation gaps pose real risks, including the propagation of inaccurate species descriptions, misclassifications of traits, and loss of trust in automated systems. By pairing automatic metrics with domain-expert assessment, our aim is to identify evaluation failures that have ethical implications for the deployment of LLMs in high-stakes QA tasks.

5 Results and Analysis

5.1 Quantitative Overview of Expert Ratings

We first report the average scores assigned by the domain expert across the five evaluation dimensions. As shown in Table 2, the model achieves high average ratings in Accuracy (4.74), Botanical Terminology Usage (4.78), and Completeness (4.53), with slightly lower but still strong scores for Relevance and Fluency (both at 4.48).

To assess whether surface-level qualities such as fluency are indicative of factual correctness, we computed the Pearson correlation coefficients between the expert-rated dimensions (Table 3). Accuracy and Completeness show a moderate correlation ($r = 0.52$), while Fluency correlates only weakly with Accuracy ($r = 0.35$) and even less with Completeness ($r = 0.17$). The weakest correlation is between Relevance and Terminology ($r = 0.08$), and Fluency shows only a modest link to Botanical Terminology, yet the highest compared to other expert-based metrics ($r = 0.42$). These findings suggest that well-written outputs are not reliable indicators of factual quality.

Metric / Expert Dimension	Mean Score
BLEU	51.48
ROUGE-L	77.13
Exact Match (EM)	0.22
BERTScore F1	0.93
Expert Accuracy	4.74
Expert Completeness	4.53
Expert Relevance	4.48
Expert Fluency	4.48
Botanical Terminology Usage	4.78

Table 2: Comparison of average automatic metric scores with expert evaluation ratings (scale: BLEU/ROUGE in %, EM in [0–1], Experts in [1–5]).

	Acc.	Comp.	Rel.	Flu.	Term.
Accuracy	1.00	0.52	0.52	0.35	0.16
Completeness		1.00	0.30	0.17	0.12
Relevance			1.00	0.17	0.08
Fluency				1.00	0.42
Terminology					1.00

Table 3: Pearson correlation matrix between expert evaluation dimensions.

5.2 Automatic Metrics vs. Expert Ratings

To evaluate how well commonly used automatic metrics align with expert human judgment, we compared BLEU, ROUGE-L, Exact Match (EM), and BERTScore against five expert-evaluated dimensions: Accuracy, Completeness, Relevance, Fluency, and Botanical Terminology Usage. Table 2 reports the mean scores across 100 evaluated samples.

Context: Twining herbaceous plant with alternate paripinnate leaves. Rachis 7 to 9 cm long, bearing 10 to 13 pairs of rectangular leaflets 15 to 20 mm long, 6-8 mm wide, decreasing in size towards the base: apex and base almost truncated. Pairs of leaflets spaced 4-7 mm apart. Upper side of the blade glabrous, underside with fairly dense semi-erect white hairs. Petiole 2 to 15 mm long before the first pair of leaflets; rachis with fairly dense semi-erect red hairs: stems with the same pubescence. **Question:** *What is the color of the hairs on the underside of the leaflets?* **Fine-tuned LLM Answer:** "The underside of the leaflets is covered with semi-erect, fairly dense white hairs." **Reference Answer:** The hairs are white. **Expert Evaluation:** Accuracy: 5, Completeness: 5, Fluency: 5 **Metrics:** BLEU: 1.7, EM: 0, ROUGE-L: 22.22, BERTScore: 0.80 *Comment: Correct and informative answer; metrics penalized due to elaboration.*
Context: Small tree, 10 to 15 m tall, or a sarmentose-lianaceous shrub with alternate leaves. Obovate leaf blade, 8 to 18 cm long, or more, 5 to 9 cm wide at the top. Wedge-shaped base, broadly rounded top, sometimes truncated, or with a short, sharp point in the middle. 7 to 9 lateral veins and, between them, nervils spaced 3-4 mm apart, parallel to each other. **Question:** *What is the base of the leaves like?* **Fine-tuned LLM Answer:** "The base of the leaves is wedge-shaped." **Reference Answer:** The base of the leaves is wedge-shaped, the apex is broadly rounded or sometimes truncate, or with a short abrupt point in the middle. **Expert Evaluation:** Accuracy: 5, Completeness: 5, Fluency: 4 **Metrics:** BLEU: 8.23, EM: 0, ROUGE-L: 46.67, BERTScore: 0.84 *Comment: Accurate and concise, but metrics suffer due to omission of less relevant details.*
Context: Petiole 1 to 5 cm long, covered with the same whitish pubescence, as well as the stem and the 2 filiform stipules 3 to 6 mm long. Axillary flowers, isolated, or in pairs, at the top of a peduncle 1 to 2 cm long. Golden yellow corolla 20 mm wide, with 5 obtriangular petals, with a broad and emarginate apex. In the center, a short column bearing numerous stamens. Calyx with 5 fairly wide triangular teeth at the base. **Question:** *Which organs are mentioned in the plant description?* **Fine-tuned LLM Answer:** "The mentioned organs are the petiole, stem, stipules, flowers, corolla, petals, column, stamens, and calyx." **Reference Answer:** The mentioned organs are the petiole, stem, stipules, corolla, column, stamens, calyx, and peduncle. **Expert:** Accuracy: 5, Completeness: 4, Fluency: 5 **Metrics:** BLEU: 50.59, EM: 0, ROUGE-L: 80.0, BERTScore: 0.96 *Comment: Nearly perfect, missing only the peduncle; EM fails despite strong factual match.*

Table 4: Examples showing mismatch between automatic metrics and expert evaluation.

Discrepancies Between Metrics and Expert Judgment: As illustrated in Table 4, even complete and accurate factual answers can receive low automatic scores due to linguistic variation or partial overlap with reference. Despite high expert ratings for accuracy and terminology usage, Exact Match (EM) remains extremely low at 0.22. This illustrates the inadequacy of EM in settings where paraphrasing and linguistic variation are common. Similarly, BLEU (51.48) and ROUGE-L (77.13) (Table 2) reflect moderate overlap but remain insensitive to omissions or hallucinations, two critical failure modes in scientific QA.

Semantic vs. Factual Fidelity: BERTScore F1 (0.93) more closely tracks expert evaluations, suggesting better alignment with semantic content. However, BERTScore cannot distinguish between correct information and plausible-sounding hallucinations, nor does it penalize factual incompleteness. These results reinforce the notion that semantic similarity does not imply factual fidelity.

Ethical Implications: These discrepancies raise serious ethical concerns. In high-stakes domains like biodiversity, law, and medicine, models can receive strong automatic scores while omitting crucial details or introducing unverifiable content. Therefore, over-reliance on surface-level metrics can mislead downstream users, researchers, or policymakers into trusting outputs that lack scientific rigor.

We provide empirical evidence for the core claim of this paper: that standard metrics such as BLEU, ROUGE, EM, and BERTScore fail to capture the factual quality of LLM-generated answers in domain-specific settings. We argue for incorporating expert validation and task-specific evaluation frameworks as ethical imperatives in future work on domain-adapted QA systems.

5.3 Alignment of Automatic Metrics with Expert Ratings

To further quantify how well automatic metrics track expert judgment, we computed Pearson correlations between BLEU, ROUGE-L, EM, and

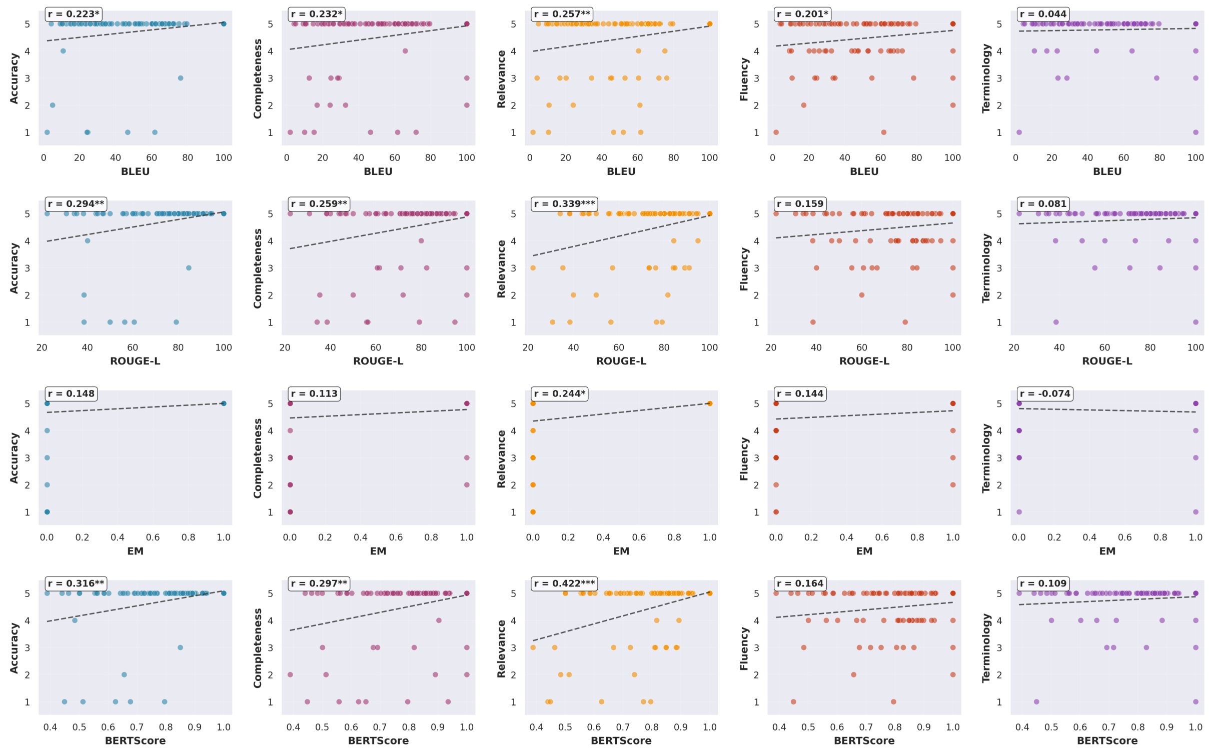

Figure 1: Scatter plot matrix showing automatic metrics vs. expert evaluation dimensions (1–5 scale). Pearson correlations shown with significance levels (*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$). n=100.

BERTScore and the five expert-rated dimensions (Accuracy, Completeness, Relevance, Fluency, Terminology). Figure 2 visualizes the results.

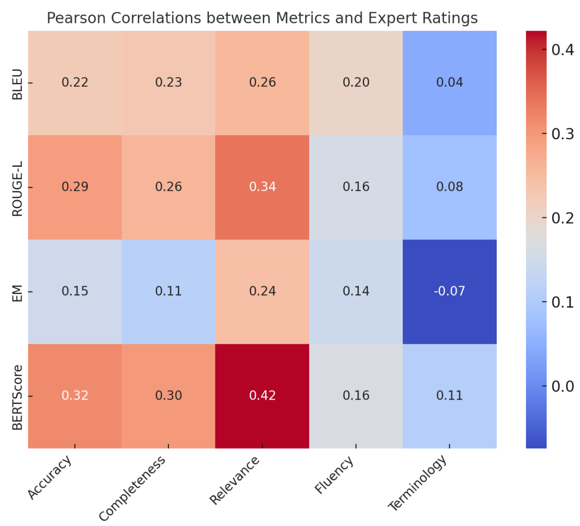

Figure 2: Pearson correlations between automatic metrics (rows) and expert-rated dimensions (columns). Values indicate weak-to-moderate alignment at best; EM is largely uninformative, and BERTScore correlates most with Relevance rather than factual Accuracy.

Overall, alignments are weak to moderate. The strongest association is BERTScore with Relevance ($r \approx 0.42$), followed by ROUGE-L with Relevance ($r \approx 0.34$). Correlations with Accuracy are only modest across metrics (BLEU ≈ 0.22, ROUGE-L ≈ 0.29, BERTScore ≈ 0.32), and associations with Fluency are uniformly low ($r \leq 0.20$). Terminology exhibits the weakest alignment overall (e.g., EM ≈ -0.07). Notably, Exact Match is effectively uncorrelated with all expert dimensions, showing its weakness when faced with paraphrased or partially correct answers. These findings reinforce that surface-similarity metrics (BLEU, ROUGE. EM) and even semantic similarity (BERTScore) do not reliably capture factual correctness, completeness, or terminological precision in domain-specific QA.

Furthermore, the correlation analysis (Figure 1) reveals significant limitations in current automatic evaluation metrics for botanical QA assessment. BERTScore demonstrates the strongest alignment with expert judgments, showing moderate correlations with semantic dimensions: Relevance ($r = 0.422^{***}$), Accuracy ($r = 0.316^{**}$), and Completeness ($r = 0.297^{**}$). ROUGE-L exhibits weaker but statistically significant correlations, particularly with Relevance ($r = 0.339^{***}$) and Completeness ($r = 0.259^{**}$). BLEU shows minimal correlations across all dimensions ($r \leq 0.257$), with only Relevance reaching significance ($r = 0.257^{**}$).

Exact Match proves largely uninformative with weak correlations ($r \leq 0.244$) and limited significance. Critically, all automatic metrics show negligible correlations with Terminology assessment ($r \leq 0.109$, mostly non-significant), which highlights their inability to capture domain-specific linguistic accuracy crucial for specialized QA systems. The moderate correlations overall (highest $r = 0.422$) indicate that automatic metrics capture only partial aspects of expert-valued quality, with BERTScore being the most reliable predictor, while human evaluation remains essential for comprehensive assessment in specialized domains.

6 Conclusion

We highlight the limitations of widely used automatic evaluation metrics: BLEU, ROUGE, Exact Match, and BERTScore in capturing the factual accuracy, completeness, and domain-specific fidelity of LLM-generated answers in scientific question answering. Our comparative analysis against expert ratings reveals that these metrics often reward superficial overlap while failing to penalize critical omissions, hallucinations, or terminological imprecision.

We argue that relying solely on these metrics can lead to misleading conclusions about model performance, particularly in high-stakes fields such as biodiversity. As illustrated through both aggregate scores and specific examples, expert-based evaluation provides a more reliable lens for assessing output quality in domain-adapted QA systems.

Future work should prioritize the development of evaluation frameworks that integrate domain expertise, task-specific criteria, and human-in-the-loop feedback. Doing so is not only methodologically sound but ethically necessary to ensure the safe deployment of LLMs in scientific and ecological applications.

7 Limitations

While our analysis highlights important shortcomings of automatic evaluation metrics in domain-specific QA, several limitations remain.

First, our study focuses on a single domain, botanical and ecological question answering using a dataset of 100 expert-rated examples. Although the findings are indicative, they may not fully generalize, to the same degree, to all other scientific or technical fields with different terminological structures or reasoning demands.

Second, expert evaluation, while more reliable than surface-level metrics, introduces its own subjectivity. Although we employed a biodiversity expert with domain knowledge, future work should include multiple annotators to assess inter-annotator agreement.

Third, our evaluation primarily addresses short-form, extractive QA responses. Longer, multi-step, or generative answers may pose different challenges, particularly around discourse coherence, reasoning chains, and multi-document grounding areas not fully captured in our current setup.

Finally, we did not explore recent or emerging evaluation methods such as LLM-as-a-judge or retrieval-augmented verification, which may complement expert-based evaluation or improve factuality assessment in future iterations.

Addressing these limitations in future work will be critical to building more robust and trustworthy evaluation pipelines for domain-adapted QA systems.

Acknowledgments

This work is co-funded by the European Union's Horizon Europe research and innovation program Cofund SOUND.AI under the Marie Sklodowska-Curie Grant Agreement No 101081674. This research is also part of the e-COL+ project (ANR-21-ESRE-0053).

References

Youssef Al Mouatamid, Jihad Zahir, Marie Bonnin, and Hajar Mousannif. 2023. *Assessing Ocean's Legal Protection Using AI: A New Dataset and a BERT-Based Classifier*.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2024a. L-eval: Instituting standardized evaluation for long context language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14388–14411, Bangkok, Thailand. Association for Computational Linguistics.

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. 2024b. Learning from mistakes makes llm better reasoner.

Jean-Christophe Bélisle-Pipon. 2024. Why we need to be careful with llms in medicine. *Frontiers in Medicine*, Volume 11 - 2024.

Andry Castro, João Pinto, Luís Reino, Pavel Pipek, and César Capinha. 2024. Large language models overcome the challenges of unstructured text data in ecology. *Ecological Informatics*, 82:102742.

X. Chen, L. Wang, M. You, W. Liu, Y. Fu, J. Xu, S. Zhang, G. Chen, K. Li, and J. Li. 2024. Evaluating and enhancing large language models' performance in domain-specific medicine: Development and usability study with docoa. *Journal of Medical Internet Research*, 26:e58158.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences.

Hao Cui, Zahra Shamsi, Gowoon Cheon, Xuejian Ma, Shutong Li, Maria Tikhanovskaya, Peter Norgaard, Nayantara Mudur, Martyna Plomecka, Paul Raccuglia, Yasaman Bahri, Victor V. Albert, Pranesh Srinivasan, Haining Pan, Philippe Faist, Brian Rohr, Ekin Dogus Cubuk, Muratahan Aykol, Amil Merchant, Michael J. Statt, Dan Morris, Drew Purves, Elise Kleeman, Ruth Alcantara, Matthew Abraham, Muqthar Mohammad, Ean Phing VanLee, Chenfei Jiang, Elizabeth Dorfman, Eun-Ah Kim, Michael P Brenner, Viren Jain, Sameera Ponda, and Subhashini Venugopalan. 2025. Curie: Evaluating llms on multitask scientific long context understanding and reasoning.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S. Rosen, Gerbrand Ceder, Kristin A. Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418.

Filip Dorm, Joseph Millard, Drew Purves, Michael Harfoot, and Oisin Mac Aodha. 2025. Large language models possess some ecological knowledge, but how much? *bioRxiv*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2025. Length-controlled alpacaeval: A simple way to debias automatic evaluators.

Alexander R. Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. Qafacteval: Improved qa-based factual consistency evaluation for summarization.

Md Meftahul Ferdaus, Mahdi Abdelguerfi, Elias Ioup, Kendall N. Niles, Ken Pathak, and Steven Sloan. 2024. Towards trustworthy ai: A review of ethical and robust large language models.

Riccardo Giorgino, Mario Alessandri-Bonetti, Andrea Luca, Filippo Migliorini, Nicolò Rossi, Giuseppe M. Peretti, and Laura Mangiavini. 2023. Chatgpt in orthopedics: a narrative review exploring the potential of artificial intelligence in orthopedic practice. *Frontiers in Surgery*, Volume 10 - 2023.

Andrew V. Gougherty and Hannah L. Clipp. 2024. Testing the reliability of an ai-based large language model to extract ecological information from the scientific literature. *npj Biodiversity*, 3(1):13.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. Knowledge-centric hallucination detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6953–6975, Miami, Florida, USA. Association for Computational Linguistics.

Jinhyuk Lee, Mujeen Sung, Jaewoo Kang, and Danqi Chen. 2021. Learning dense representations of phrases at scale.

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

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

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2022. A token-level reference-free hallucination detection benchmark for free-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6723–6737, Dublin, Ireland. Association for Computational Linguistics.

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 in Bioinformatics*, 23(6):bbac409.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar. 2023. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265.

Ayoub Nainia, Régine Vignes-Lebbe, Eric Chenin, Maya Sahraoui, Hajar Mousannif, and Jihad Zahir. 2024. Floraner: A new dataset for species and morphological terms named entity recognition in french botanical text. *Data in Brief*, 56:110824.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016a. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016b. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Ravi Shanker Raju, Swayambhoo Jain, Bo Li, Jonathan Lingjie Li, and Urmish Thakker. 2024. Constructing domain-specific evaluation sets for LLM-as-a-judge. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 167–181, Miami, Florida, USA. Association for Computational Linguistics.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large language models encode clinical knowledge.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with large language models.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022a. Galactica: A large language model for science.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022b. Galactica: A large language model for science.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.