Who is the richest club in the championship? Detecting and Rewriting Underspecified Questions Improve QA Performance

Yunchong Huang
ILLC, University of Amsterdam
franzhuang027@gmail.com

Gianni Barlacchi
Amazon AGI *
gbarlac@amazon.com

Sandro Pezzelle
ILLC, University of Amsterdam
s.pezzelle@uva.nl

Abstract

Large language models (LLMs) perform well on well-posed questions, yet standard question-answering (QA) benchmarks remain far from solved. We argue that this gap is partly due to *underspecified questions*—queries whose interpretation cannot be uniquely determined without additional context. To test this hypothesis, we introduce an LLM-based classifier to identify underspecified questions and apply it to several widely used QA datasets, finding that 16% to over 50% of benchmark questions are underspecified and that LLMs perform significantly worse on them. To isolate the effect of underspecification, we conduct a controlled rewriting experiment that serves as an upper-bound analysis, rewriting underspecified questions into fully specified variants while holding gold answers fixed. QA performance consistently improves under this setting, indicating that many apparent QA failures stem from question underspecification rather than model limitations. Our findings highlight underspecification as an important confound in QA evaluation and motivate greater attention to question clarity in benchmark design.

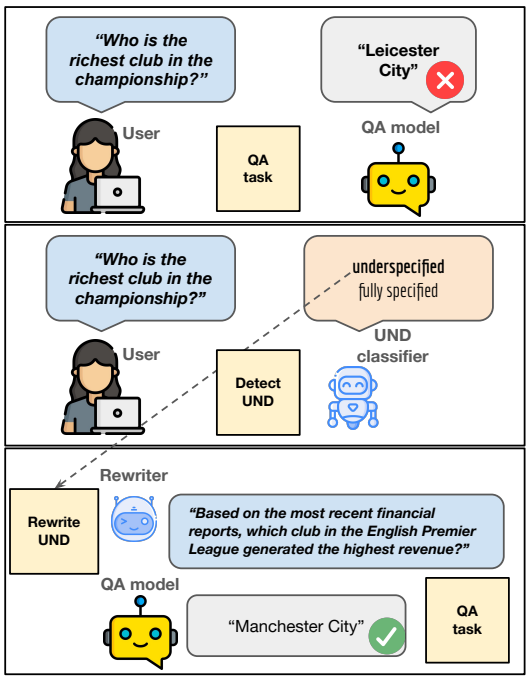

Figure 1: Top: One real question from Natural Questions (NQ) dataset (Kwiatkowski et al., 2019) and corresponding *wrong* answer by a QA model, i.e., `GPT-4o` (OpenAI et al., 2024). Middle: We build an LLM-based classifier to detect underspecified (UND) questions in QA benchmarks. Bottom: We turn UND questions into fully specified ones using an LLM-based rewriter and verify that, by removing underspecification, performance on the QA task significantly improves.

1 Introduction

Large Language Models (LLMs) perform well on clearly specified factual queries, yet widely used question-answering (QA) benchmarks remain unsolved (Kwiatkowski et al., 2019; Yang et al., 2018; Joshi et al., 2017; Krishna et al., 2025; Sorodoc et al., 2025). In parallel, a growing body of work shows that LLMs struggle with vague, ambiguous, or incomplete queries (Tanjim et al., 2025a; Liu et al., 2023; Zhang et al., 2024; Zhang and Choi, 2025; Tanjim et al., 2025b; Qian et al., 2024; Herlihy et al., 2024).

In this paper, we argue that these two observations are closely related. Specifically, we hypothesize that a non-trivial portion of the apparent difficulty of QA benchmarks stems not from model limitations alone, but from the presence of underspecified questions, that is, queries whose interpretation cannot be uniquely determined without additional contextual information, which is unavailable at evaluation time.

The notion of underspecification originates in linguistic theory, where it describes expressions that admit multiple possible meanings and require additional contextual information to be resolved (Egg, 2010; Sennet, 2023; Nieuwland and Van Berkum, 2008; Sorensen, 2022; van Rooij, 2011; Kennedy, 2011; Carston, 2002; Belleri, 2014; Grice, 1957).

*Work done outside Amazon.

Findings of the Association for Computational Linguistics: ACL 2026, pages 26266–26281
July 2-7, 2026 ©2026 Association for Computational Linguistics

An underspecified utterance does not fully encode its intended meaning and relies on shared common ground, discourse context, or pragmatic inference to narrow down its interpretation (Frisson, 2009; Harris, 2020; Pezzelle, 2023; Wildenburg et al., 2024; Bach, 2004; Stalnaker, 2002). While such mechanisms are often available in natural communication, in the context of QA benchmarks such information is typically unavailable, making it impossible to recover the intended answer from the question alone uniquely. The top panel of Fig. 1 illustrates this issue with an example from the Natural Questions (NQ) dataset (Kwiatkowski et al., 2019)—*Who is the richest club in the championship?*—where the specific referent of *championship*, the definition of *richest*, and the specific time frame to answer the question are all left underspecified.

We hypothesize that underspecified questions are prevalent in existing QA benchmarks and systematically reduce reported model performance. To test this hypothesis, we first introduce an LLM-based classifier that distinguishes between fully specified (FS) and underspecified (UND) questions. Applying it to several widely used QA datasets, we find that all benchmarks contain a substantial proportion of underspecified questions, ranging from around 16% of total questions in TriviaQA (Joshi et al., 2017) to over 50% in FRAMES (Krishna et al., 2025). We then show that state-of-the-art LLMs perform significantly worse on UND questions than on FS ones across datasets. To further isolate the effect of underspecification, we conduct a controlled rewriting experiment that serves as an upper-bound analysis. By rewriting underspecified questions into fully specified variants while holding the gold answer fixed, we evaluate how QA models perform when underspecification is removed, and find that QA performance consistently improves on rewritten questions, indicating that many errors stem from question formulation rather than model limitations.

Our contributions are as follows: (1) we provide a systematic analysis of underspecification in widely used QA benchmarks, showing that it is both prevalent and impactful to evaluation reliability; (2) we introduce an automated method for identifying underspecified questions; (3) we demonstrate, via controlled rewriting experiments, that many apparent QA failures disappear once underspecification is removed; and (4) we release our data and code publicly to support reproducibility and future research at https://github.com/franzyellow/Underspecification-QA-conf-paper.

2 Approach

Step 1: Detecting underspecified questions We first develop an LLM-based classifier (the *UND classifier*) to distinguish fully specified (FS) from underspecified (UND) queries, and use it to extract FS and UND subsets from standard QA benchmarks (see middle panel in Fig. 1). To ensure reliability, we select the best-performing model from a suite of SotA open-weight LLMs (e.g., Qwen3, DeepSeek R1) by evaluating them on existing underspecification-annotated data. We further strengthen this classification by validating the model against a smaller, expert-annotated gold standard subset.

Step 2: Assessing LLM performance on QA With the benchmarks partitioned, we evaluate two state-of-the-art proprietary LLMs (the *QA models*) on four diverse QA benchmarks. This step is a comparative analysis designed to test our core hypothesis: that standard accuracy metrics are systematically diminished by the UND portion of the data.

Step 3: Rewriting UND questions To confirm that lower performance on UND questions is due to missing context rather than a lack of model knowledge, we conduct a controlled rewriting experiment. We use an LLM-based rewriter to resolve the ambiguities identified in Step 1. Critically, this rewriter has access to the ground-truth (gold) answers and the classifier's reasoning. This "oracle" setup serves as a controlled intervention to transform UND questions into FS variants while holding the intended answer fixed.

Step 4: Reassessing LLM's QA performance In this final step, we reassess the performance of the *QA models* from Step 2 on the rewritten UND questions and compare it to their performance on the original UND queries, evaluating whether rewriting leads to significant improvements. This step establishes an upper bound of performance, demonstrating how much of the failure disappears when the benchmark's underspecified elements are removed.

3 Experiments

3.1 Step 1: Detecting underspecified questions

Data To build a reliable *UND classifier*, we curated UNDER, a multi-source dataset with FS/UND labels and comprising 855 questions from CLAMBER (Zhang et al., 2024), IN3 (Qian et al., 2024), and CoCoNot (Brahman et al., 2024).[1] UNDER contains 855 questions in total, i.e., 431 FS (50.4%), and 424 UND questions (49.6%) encompassing different types of underspecification. These questions are annotated with different terminologies in corresponding source datasets, but have been mapped to the unified UND/FS binary labeling in our configuration.

To ensure linguistic rigor, we further developed UNDER-gold, a subset of 150 instances from the same datasets and from uncontroversially annotated examples (verified by the authors) in AmbigNQ (Min et al., 2020), with no overlap with UNDER. Each QA pair was manually reviewed by an author with expertise in formal linguistics to verify the alignment between automated labels and theoretical underspecification. The manual annotation followed our working taxonomy of underspecification showcased in Table 1[2], ensuring that each UND/FS label is grounded in principled reasoning and expert-verified and that all UND questions are identified with a finer-grained taxonomical type.

Models We experiment with `Qwen3` with four different parameter sizes (4B, 8B, 14B, and 32B) in the "thinking mode" (Yang et al., 2025); `DeepSeek R1` (Guo et al., 2025) distilled models with four parameter sizes (1.5B, 7B, 14B, and 32B); `Llama-3.2-3B-Instruct` (Meta, 2024) and `Llama-3.3-70B-Instruct` (Meta AI, 2024).

Experimental setup We prompt the LLMs via instructions to act as expert analysts and classify input questions as either FS or UND (UND covers various types of underspecification). The prompt includes a task description, the target question, and an explicit requirement for the output format. [3] We analyze task accuracies and macro F1 values of all models on both UNDER and UNDER-gold to select the best-performing model.

UND types	Example	Source
Type 1: Missing necessary components	*Ok Google, what's the capital?* Which country/region is being asked about?	CoCoNot
Type 2: Undetermined lexicons or references	*When was the last time the Giants went to the playoffs?* "The Giants": the football team or the baseball team?	CLAMBER
Type 3: Undetermined perspective or granularity	*When was the First World War broke out?* Is "broke out" in the political sense or in the military sense?	AmbigNQ
Type 4: Undetermined standard or preference	*Recommend the best smartwatches available in 2023.* What's the specific standard of being "best'"?	IN3

Table 1: The taxonomic categories of underspecified queries in QA, with examples from multiple annotated datasets related to semantic underspecification.

Results While several models perform comparably on the UND/FS classification task (see Table 4 in the Appendix C), `Qwen3-4B` achieves the highest overall accuracy (0.71) and macro F1 (0.70) on the UNDER dataset. Its performance of 0.77 accuracy and 0.76 macro F1 on UNDER-gold also makes it the best among all tested models. Across both datasets, `Qwen3-4B` yields the highest average accuracy (0.74) and average macro F1 (0.73), outperforming the second best model `Qwen3-32B` (average accuracy 0.71, average macro F1 0.71) by a notable margin. Based on these results, we selected this model as the UND classifier for all subsequent experiments. These results indicate that this classifier satisfactorily distinguishes UND from FS questions, especially when labels align closely with expert annotations (i.e., UNDER-gold).

3.2 Step 2: Assessing LLM QA performance

Data We experiment with four widely used QA datasets without underspecification-related annotations: Natural Questions (NQ) (Kwiatkowski et al., 2019), HotpotQA (Yang et al., 2018), TriviaQA (Joshi et al., 2017), FRAMES (Krishna et al., 2025). In this experiment, we consider only the questions themselves, without providing any context passage to the model (e.g., the context paragraphs in HotpotQA). For FRAMES, we include all 824 ⟨question, annotated answer⟩ pairs. From

[1] We did not consider the popular AmbigNQ dataset (Min et al., 2020) for two reasons: (i) its bottom-up crowdsourcing procedure introduced substantial cross-annotation inconsistencies on the perceptive threshold of underspecification; (ii) annotators were instructed to enumerate all plausible answers using Wikipedia, resulting in a much finer-grained and knowledge-intensive notion of underspecification. Thus, automatically mapping AmbigNQ instances to binary UND/FS labels without heavy manual verification is unreliable and, therefore, incompatible with the procedure used for the other datasets. However, we did consider AmbigNQ to construct a small, hand-curated "gold" test set, as explained below.

[2] See Appendix A for a full illustration of the taxonomy.

[3] Refer to Appendix B for the specific prompt used.

Models/Datasets	NQ UND → NQ Rewr	HotpotQA UND → HotpotQA Rewr	TriviaQA UND → TriviaQA Rewr	FRAMES UND → FRAMES Rewr
GPT-4o	37.0% → 57.3% (+20.3%)	34.6% → 51.8% (+17.2%)	75.8% → 83.6% (+7.8%)	24.4% → 41.6% (+17.2%)
Gemini-2.5-Flash	38.8% → 50.0% (+11.2%)	41.2% → 50.6% (+9.4%)	76.0% → 74.4% (-1.6%)	37.1% → 46.5% (+9.4%)

Table 2: Comparison of F1 score for `GPT-4o` and `Gemini-2.5-Flash` across datasets. UND → Rewritten (+Δ).

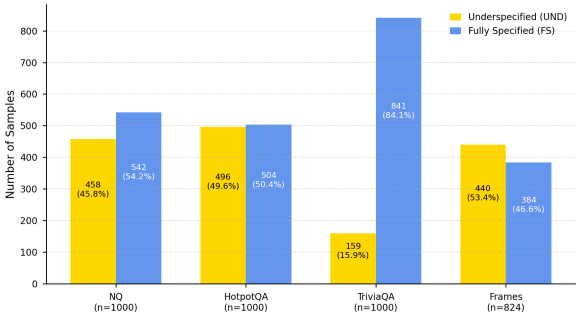

Figure 2: Proportion of FS/UND questions in each of the four source datasets included in QA-ensemble.

each of the remaining datasets, we uniformly sample 1,000 instances, yielding a total of 3,824 data points. We refer to this aggregated collection as **QA-ensemble**.

Models We experiment with two proprietary LLMs, `GPT-4o` (`gpt-4o-2024-11-20`) (OpenAI, 2024) and `Gemini-2.5-Flash` (Comanici et al., 2025), which we use as our *QA models*.

Experimental setup We first apply the *UND classifier* to QA-ensemble to automatically label questions as fully specified (FS) or underspecified (UND). The *QA models* are then prompted to answer all questions, and performance is evaluated separately on the UND and FS subsets.[4] We assess performance using both a standard token-level F1 score and an LLM-as-judge metric, namely the Nvidia Answer Accuracy (AA) within the RAGAS framework (ExplodingGradients, 2024, 2025). [5] For each model, we conduct independent t-tests to determine whether performance on UND questions is significantly lower than on FS questions.

Results. Fig. 2 presents the FS/UND classification results of samples from the four source QA datasets provided by *UND classifier*. The proportion of UND queries varies between datasets, with the lowest observed in TriviaQA (15.9%) and the highest observed in FRAMES (53.4%). These results indicate a widespread presence of underspecified queries in all source QA datasets. By running *QA models* on these partitions, we report a statistically significantly lower performance on UND compared to FS questions for all models and all datasets in QA-ensemble (see Appendix D.1 for detailed results).

3.3 Step 3: Rewriting UND questions

Data and models We use all UND questions from QA-ensemble. We experiment with `GPT-4o` and `Gemini-2.5-Flash` as *Rewriter models*.

Experimental setup We prompt the *Rewriter models* to rewrite UND questions. The prompt includes golden answers to these UND queries annotated in their original QA datasets and the classification reasoning provided by the *UND classifier* in Step 2 to supervise the rewriting task.[6] Then, we run the *QA models* on these rewritten UND questions and report the percentage of queries that are classified as FS after the rewriting.

Results The majority of rewritten questions are classified as fully specified, with proportions ranging from 64% to over 86%. This indicates that the rewriting step effectively removes underspecification, converting UND questions into FS ones. In the Appendix C, Table 5 reports the *UND classifier*'s predictions on questions rewritten by our two *Rewriter models*, `GPT-4o` and `Gemini-2.5-Flash`.

3.4 Step 4: Reassessing QA performance

Experimental Setup We reassess *QA models* on the rewritten queries from Step 3 using the same QA prompt as in Step 2. To avoid bias from using the same LLM for both rewriting and answering, we adopt a cross-assignment setup: queries rewritten by `GPT-4o` are answered by `Gemini-2.5-Flash`, and *vice versa*. We then compare QA performance on rewritten versus original UND queries using independent t-tests.

Results As shown in Table 2, `GPT-4o` and `Gemini-2.5-Flash` achieve broadly similar performance across all datasets. Rewriting queries to make them fully specified consistently improves results with the only exception of TriviaQA. This

[4] See Appendix B for the full QA prompt.

[5] Nvidia AA measures alignment between a generated answer and a reference by averaging two independent LLM judgments on a discrete scale (0, 2, 4), normalized to [0, 1].

[6] Please refer to Appendix B for the rewriting prompt.

Question	Source	QA Model	Golden Answers	Model Answer	Qwen3-4B Pred	Components leading to UND (Qwen3-4B)
Who is the richest club in the championship?	NQ	GPT-4o	['Manchester City']	['Leicester City']	UND	Underspecified standard: 'richest'; undetermined reference: 'Championship'
What was the code name of the landing barge primarily used to provide hot meals to the landing crew?	HotpotQA	GPT-4o	['Operation Neptune']	['HMS Menestheus']	UND	Potential references: a historically documented vessel, a fictional construct, or a specific operation?
Who wrote the piece of music recognised in much of the Western world as "The Wedding March" (or as "Here Comes The Bride")?	TriviaQA	Gemini-2.5-Flash	['Felix Mendelssohn', 'mendelsson bartholdy' ...]	['Richard Wagner']	UND	Potential references: 1) 'The Wedding March' from Wagner's *Lohengrin*, 2) 'Here Comes the Bride' from the 1935 film *The Wizard of Oz*, 3) 'Wedding March' from Mendelssohn's *A Midsummer Night's Dream*.
Who developed the first effective vaccine against the disease that killed the father of a famous Hungarian composer born in 1811?	FRAMES	Gemini-2.5-Flash	[Almroth Edward Wright]	['Edward Jenner']	UND	The specific disease causing the death remains ambiguous. Historical accounts suggest Franz Liszt Sr. died of typhoid fever, but this is contested.

Table 3: Examples of UND queries with incorrect answers by *QA models*

is expected since the dataset is explicitly designed to contain quiz questions that already provide sufficient information for answering. For example, `GPT-4o` on HotpotQA rises from 34.6% to 51.8% (+17.2%), and `Gemini-2.5-Flash` on NQ rises from 38.8% to 50.0% (+11.2%). These improvements indicate that when questions are well-formed and include all critical information, models can answer correctly, suggesting that the main source of errors in the UND setting is incomplete or unclear question formulation rather than limitations of the QA systems themselves. More detailed results with both F1 and Nvidia AA metrics are reported in in Appendix C, Table 6. Additionally, to rule out the potential effect of knowledge leakage where rewritten questions may contain vocabulary directly from golden answers, allowing models to exploit surface-level cues, we conduct a quantitative analysis to verify the lexical overlap between questions (original and rewritten) and their golden answers is minor. To do so, we use Jaccard similarity and n-gram overlap (unigram and bigram F1) as metrics and report the results in the Appendix C, Table 7. This analysis shows that, while rewritten questions often exhibit statistically significantly higher lexical overlap with golden answers than original ones, absolute overlap values remain low (< 0.1) across all metrics and configurations. This suggests that lexical similarities between rewritten questions and golden answers are marginal and unlikely to account for the observed performance gains.

Analysis We conduct a manual qualitative analysis of UND queries for which *QA models* produce incorrect answers according to both F1 and NVIDIA AA metrics,[7] with some examples shown in Table 3. We qualitatively observe that, in these cases, *QA models* often generate answers that conflict with the gold annotation in the dataset, while this mismatch is resolved after rewriting. Additional examples are reported in Appendix E, Table 8 and Table 9.

4 Conclusion

Our study highlights the widespread presence of underspecified questions in current QA benchmarks, which can affect the reliability of QA model evaluation. We introduce an LLM-based classifier to detect such queries and a rewriter to transform them into fully specified forms. Across multiple widely used QA datasets, LLMs consistently struggle with underspecified questions when no supplementary contextual information (e.g., contextual paragraph) is provided, but performance improves significantly once underspecification/ambiguities are resolved, showing that many apparent QA failures reflect question clarity rather than model limitations.

These findings highlight underspecification as a universally critical factor in QA evaluation, and we argue that future work should pay extra attention to QA underspecification by identifying, annotating, and possibly rewriting these questions. This will help shed light on the real capabilities of LLMs as effective and reliable QA models.

Limitations

We acknowledge several limitations in this work. First, we adopted an off-the-shelf LLM as the UND-classifier. While its performance is reasonably strong, it is not optimal, and there remains room to improve classification accuracy and macro-F1 through prompt-based optimization or knowledge enhancement (e.g., RAG-based methods).

[7] We consider samples with values <0.5 on both metrics.

Second, implementation of our pipeline could be further improved by incorporating recent advances in LLM engineering, such as automatic prompt optimization and agentic frameworks.

Third, due to limited time and human resources, certain human expert verifications were not thoroughly conducted. For example, in Step 1, more manual inspection of the automatic mapping from original annotations to UND/FS labels could help mitigate cross-dataset inconsistencies in underspecification standards; in Step 3, comprehensive human review of rewritten questions could improve rewriting quality, recover failed cases, and prevent successfully rewritten questions from becoming overly trivial for downstream QA models.

Despite these limitations, we believe that our work provides a systematic and reproducible pipeline for detecting and rewriting underspecified QA queries, and that the main conclusions of our study remain robust.

Acknowledgments

We would like to thank Jelke Bloem, Martha Lewis, and Malvin Gattinger for providing feedback on YH's master's thesis, which laid the foundations of this work. We are also grateful to the members of the *Dialogue Modelling Group (DMG)* and the *Multimodality, Language, and Interpretability (Mulini) lab* at the University of Amsterdam for their insightful feedback on the manuscript. We also acknowledge the use of AI assistants (Claude and ChatGPT) for code debugging in experiments.

References

Kent Bach. 2004. Context ex machina. In Zoltan Gendler Szabo, editor, *Semantics Versus Pragmatics*, pages 15–44. Oxford University Press UK.

Delia Belleri. 2014. *Semantic Under-Determinacy and Communication*. Palgrave-Macmillan, London/Basingstoke.

Faeze Brahman, Sachin Kumar, Vidhisha Balachandran, Pradeep Dasigi, Valentina Pyatkin, Abhilasha Ravichander, Sarah Wiegreffe, Nouha Dziri, Khyathi Chandu, Jack Hessel, Yulia Tsvetkov, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. 2024. The art of saying no: contextual noncompliance in language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Harry Bunt. 2007. Semantic underspecification: Which technique for what purpose? In Harry Bunt and Reinhard Muskens, editors, *Computing Meaning (Volume 3)*, pages 55–85. Springer Netherlands, Dordrecht.

Robyn Carston. 2002. *Pragmatics and Linguistic Underdeterminacy*, chapter 1. John Wiley & Sons, Ltd.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, and Henrik Jacobsson. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Markus Egg. 2010. Semantic underspecification. *Language and Linguistics Compass*, 4(3):166–181.

ExplodingGradients. 2024. Ragas: Supercharge your llm application evaluations. `https://github.com/explodinggradients/ragas`.

ExplodingGradients. 2025. Nvidia metrics. Accessed: 2025-11-18.

Steven Frisson. 2009. Semantic underspecification in language processing. *Language and Linguistics Compass*, 3:111–127.

Herbert Paul Grice. 1957. Meaning. *Philosophical Review*, 66(3):377–388.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 180 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Nature*, 645:633–638.

Daniel W. Harris. 2020. What makes human communication special? Unpublished book manuscript, CUNY Graduate Center.

Christine Herlihy, Jennifer Neville, Tobias Schnabel, and Adith Swaminathan. 2024. On overcoming miscalibrated conversational priors in llm-based chatbots. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, UAI '24. JMLR.org.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Christopher Kennedy. 2011. 23. ambiguity and vagueness: An overview. In Claudia Maienborn, Klaus von Heusinger, and Paul Portner, editors, *Volume 1*, pages 507–535. De Gruyter Mouton, Berlin, Boston.

Christopher Kennedy and Louise Mcnally. 2010. Color, context, and compositionality. *Synthese*, 174(1):79–98.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. 2025. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4745–4759, Albuquerque, New Mexico. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah Smith, and Yejin Choi. 2023. We're afraid language models aren't modeling ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 790–807, Singapore. Association for Computational Linguistics.

Meta. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. Accessed: 2025-07-13.

Meta AI. 2024. Llama 3.3-70b-instruct. Hugging Face model page. Retrieved June 17, 2025.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

Mante S. Nieuwland and Jos J. A. Van Berkum. 2008. The neurocognition of referential ambiguity in language comprehension. *Language and Linguistics Compass*, 2(4):603–630.

OpenAI. 2024. Hello gpt-4o. Accessed: 2025-07-19.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and 400 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Sandro Pezzelle. 2023. Dealing with semantic underspecification in multimodal NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12098–12112, Toronto, Canada. Association for Computational Linguistics.

Cheng Qian, Bingxiang He, Zhong Zhuang, Jia Deng, Yujia Qin, Xin Cong, Zhong Zhang, Jie Zhou, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. Tell me more! towards implicit user intention understanding of language model driven agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1088–1113, Bangkok, Thailand. Association for Computational Linguistics.

Adam Sennet. 2023. Ambiguity. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Summer 2023 edition. Metaphysics Research Lab, Stanford University.

Roy Sorensen. 2022. Vagueness. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Winter 2023 edition. Metaphysics Research Lab, Stanford University.

Ionut Teodor Sorodoc, Leonardo F. R. Ribeiro, Rexhina Blloshmi, Christopher Davis, and Adrià de Gispert. 2025. GaRAGe: A benchmark with grounding annotations for RAG evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17030–17049, Vienna, Austria. Association for Computational Linguistics.

Robert Stalnaker. 2002. Common ground. *Linguistics and Philosophy*, 25(5):701–721.

Md Mehrab Tanjim, Xiang Chen, Victor Bursztyn, Uttaran Bhattacharya, Mai Tùng, Vaishnavi Muppala, Akash Maharaj, Saayan Mitra, Eunyee Koh, Yunyao Li, and Ken Russell. 2025a. Detecting ambiguities to guide query rewrite for robust conversations in enterprise ai assistants. *arXiv*.

Mehrab Tanjim, Yeonjun In, Xiang Chen, Victor Bursztyn, Ryan A. Rossi, Sungchul Kim, Guang-Jie Ren, Vaishnavi Muppala, Shun Jiang, Yongsung Kim, and Chanyoung Park. 2025b. Disambiguation in conversational question answering in the era of LLMs and agents: A survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9537–9550, Suzhou, China. Association for Computational Linguistics.

Robert van Rooij. 2011. Vagueness and linguistics. In Giuseppina Ronzitti, editor, *Vagueness: A Guide*, pages 123–170. Springer Netherlands, Dordrecht.

Frank Wildenburg, Michael Hanna, and Sandro Pezzelle. 2024. Do pre-trained language models detect and understand semantic underspecification? ask the DUST! In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9598–9613, Bangkok, Thailand. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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.

Michael JQ Zhang and Eunsol Choi. 2025. Clarify when necessary: Resolving ambiguity through interaction with LMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5526–5543, Albuquerque, New Mexico. Association for Computational Linguistics.

Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru Liang, and Tat-Seng Chua. 2024. CLAMBER: A benchmark of identifying and clarifying ambiguous information needs in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10746–10766, Bangkok, Thailand. Association for Computational Linguistics.

A A Working Taxonomy of Underspecified Queries in Question Answering (QA)

Type 1: Missing necessary components. An underspecified query under this category contains at least one expression that is missing a commonly expected component conceptually tied to it (e.g., an implicit expected argument of a predicative element). Consequently, the semantic content of the expression at issue is linguistically incomplete and undertermined, with several different interpretations possible. This category is closely related to *Missing Constituents* or *Conceptual Truncation* discussed in the framework of Linguistic (Semantic) Underdeterminacy (Carston, 2002; Belleri, 2014).

Type 2: Undetermined lexicons or references. An underspecified query under this category contains at least one expression with lexical or referential ambiguity. Multiple same-level concepts or entities can be mapped to this expression at issue to serve as potential lexical entries or referents. It is impossible to fully determine which one is intended by the user based on the provided content. This type is closely related to *lexical ambiguity* and *referential ambiguity* (Kennedy, 2011; Sennet, 2023; Nieuwland and Van Berkum, 2008), the concept of *indexical reference* in the Linguistic Underdeterminacy framework (Carston, 2002), and the *Syntactically but not semantically homogeneous ambiguities* discussed as a type of semantic underspecification in Egg (2010).

Type 3: Undetermined perspective or granularity. An underspecified query under this category contains at least one expression where the general meaning is in place, but its specific interpretation can still vary based on different heterogeneous perspectives or granularity levels adopted. Multiple interpretations of different natures or levels are plausible for such an expression, and it's impossible to fully determine which one is intended based on the provided content. As a result, the undeterminacy of this expression leads to multiple possible interpretations of the query, rendering it underspecified. Aligning with the view of Belleri (2014), we hold that it is not a type of lexical ambiguity as Kennedy (2011); Kennedy and Mcnally (2010) claims and it is also not a type of vagueness, as plausible perspectives or granularity levels are more definite and objectively acknowledged, instead of being completely a matter of contextual/subjective standard. We also claim that this category is diverging from prototypical cases of "missing constituents", as the linguistic intuition in cases under this category is more about inner interpretation of an expression at issue, instead of lacking external elements for the complete semantic saturation.

Type 4: Undetermined standard or preference. An underspecified query under this category contains at least one expression where the general meaning is in place, but its specific interpretation is vague due to unspecified contextual standards or subjective criteria. A wide range of fine-grained interpretations is possible based on contextual or subjective needs, and it's impossible to fully determine which one is intended by the user based on the provided content. As a result, the undeterminacy of this expression leads to many, or even an infinite number of, possible interpretations of the query. This category is closely related to the prototypical *vagueness* Kennedy (2011) and *semantic imprecision* proposed by Bunt (2007). In the Linguistic (Semantic) Underdeterminacy framework, phenomena discussed under *adjustments (overspecifying/underspecifying) of linguistically encoded concepts* (Carston, 2002) and *gradable expressions depending on standards or comparison classes*

(Belleri, 2014) can also be attributed to this category.

B Prompts

The prompt for *UND classifier*

<SYSTEM_PROMPT> You are an expert analyst. Your task is to analyze and determine whether an input user query is "fully specified" or "underspecified". </SYSTEM_PROMPT>

Analyze the following input user query:

{"query": "When did the nuclear accident happen?"}

Please provide your analysis in the following JSON format:

{"query": "When did the nuclear accident happen?", "reasoning": "[YOUR_DETAILED_REASONING]", "judgment": "[fully specified/underspecified]"}

The Prompt for *QA models*

<SYSTEM_PROMPT> Answer the question with a concise response. Return answers as a list of strings. If there's only one answer, return a single-item list. Each answer should be brief and direct. </SYSTEM_PROMPT>
"role": "system", "content": [SYSTEM_PROMPT],
"role": "user", "content": [QUESTION]

The prompt for *Rewriter models*

<SYSTEM_PROMPT>You are a professional question optimization expert. Please modify the underspecified question to a fully specified version based on the provided clues.
Requirements:
1. Keep the core intent of the question unchanged
2. Add necessary contextual information
3. Eliminate underspecified elements and make the question clear
4. Ensure the modified question can be directly answered with the provided short answer without dispute
Please only return the modified question, do not include any other explanations. </SYSTEM_PROMPT>
The original question: [QUESTION]
Short answer: [GOLD_ANSWER]
Reasoning: [MODEL_CLASSIFIER_REASONING]

Please analyze the underspecified elements in the original question, then modify the question to a fully specified version based on the short answer and reasoning.

C Results

Please refer to Tables 4 - 7.

D Figures

D.1 Step 2: UND-FS Performance Comparisons of Models-QA

Please refer to Fig. 3 - 6.

D.2 Step 4: Original-Rewritten Performance Comparisons of Models QA

Please refer to Fig. 7 - 10.

E More Selected Examples from Qualitative Analysis

- Please refer to Table 8 for selected examples of UND queries with incorrect answers by *QA models*.
- Please refer to Table 9 for selected examples of UND queries rewritten by *Rewriter* models, which were initially answered incorrectly by *QA models* but answered correctly in a subsequent round after rewriting.

	Qwen3-4B			Qwen3-8B			Qwen3-14B			Qwen3-32B			Llama-3.2-3B		
	UNDER	UNDER-gold	Avg	UNDER	UNDER-gold	Avg	UNDER	UNDER-gold	Avg	UNDER	UNDER-gold	Avg	UNDER	UNDER-gold	Avg
accuracy	**0.71**	**0.77**	**0.74**	0.68	0.71	0.70	0.70	0.67	0.69	0.70	0.72	0.71	0.51	0.71	0.61
macro F1	**0.70**	**0.76**	**0.73**	0.67	0.69	0.68	**0.70**	0.66	0.68	**0.70**	0.71	0.71	0.38	0.56	0.47
	DS-R1-1.5B			DS-R1-7B			DS-R1-14B			DS-R1-32B			Llama-3.3-70B		
	UNDER	UNDER-gold	Avg	UNDER	UNDER-gold	Avg	UNDER	UNDER-gold	Avg	UNDER	UNDER-gold	Avg	UNDER	UNDER-gold	Avg
accuracy	0.55	0.66	0.61	0.62	0.74	0.68	0.68	0.68	0.68	0.67	0.73	0.70	**0.71**	0.68	0.70
macro F1	0.49	0.56	0.53	0.62	0.73	0.68	0.68	0.66	0.67	0.66	0.70	0.68	**0.70**	0.68	0.69

Table 4: An overview of the performance on **UNDER** and **UNDER-gold** across the selected LLMs.

	NQ-Rewr		HotpotQA-Rewr		TriviaQA-Rewr		FRAMES-Rewr	
	GPT-4o	Gemini-2.5-Flash	GPT-4o	Gemini-2.5-Flash	GPT-4o	Gemini-2.5-Flash	GPT-4o	Gemini-2.5-Flash
FS	85.4%	84.5%	72.2%	83.6%	86.1%	83.6%	86.2%	83.6%

Table 5: For each dataset in QA-ensemble: Percentage of queries rewritten by the *Rewriter model* (either `GPT-4o` or `Gemini-2.5-Flash`) which are classified as FS by the *UND classifier*. This proportion is high: 64–86%.

	NQ		HotpotQA		TriviaQA		FRAMES	
	Orig UND	Rewr	Orig UND	Rewr	Orig UND	Rewr	Orig UND	Rewr
GPT-4o F1	37.0%	57.3%	34.6%	51.8%	75.8%	83.6%	24.4%	41.6%
GPT-4o AA	58.0%	73.4%	42.2%	61.3%	84.3%	92.8%	27.2%	51.5%
Gemini-2.5-Flash F1	38.8%	50.0%	41.2%	50.6%	76.0%	74.4%	37.1%	46.5%
Gemini-2.5-Flash AA	54.9%	68.1%	45.3%	60.5%	84.3%	87.6%	42.0%	54.6%

Table 6: The performance of QA models on original UND queries and their rewritten counterparts across QA datasets.

	Jaccard				Unigram F1				Bigram F1			
datasets + QA models	Orig	Rewr	Δ	Sig	Orig	Rewr	Δ	Sig	Orig	Rewr	Δ	Sig
NQ												
GPT-4o	0.0293	0.0477	0.0184	$p < 0.001$	0.0488	0.077	0.0282	$p < 0.001$	0.0049	0.0196	0.0148	$p < 0.001$
Gemini-2.5-Flash	0.0293	0.0711	0.0418	$p < 0.001$	0.0488	0.1074	0.0586	$p < 0.001$	0.0049	0.0512	0.0463	$p < 0.001$
HotpotQA												
GPT-4o	0.0368	0.0379	0.0011	p=0.8828	0.0606	0.0611	0.0005	p=0.7418	0.0204	0.0203	-0.0001	p=0.3560
Gemini-2.5-Flash	0.0368	0.0648	0.028	$p < 0.001$	0.0606	0.1034	0.0428	$p < 0.001$	0.0204	0.0535	0.0331	$p < 0.001$
TriviaQA												
GPT-4o	0.0203	0.0319	0.0116	$p < 0.001$	0.0254	0.0417	0.0164	$p < 0.001$	0.0038	0.0076	0.0037	p 0.01
Gemini-2.5-Flash	0.0203	0.0471	0.0267	$p < 0.001$	0.0254	0.0662	0.0409	$p < 0.001$	0.0038	0.0187	0.0149	$p < 0.001$
FRAMES												
GPT-4o	0.0506	0.0656	0.0151	$p < 0.001$	0.0725	0.0912	0.0187	$p < 0.001$	0.0312	0.0475	0.0164	$p < 0.001$
Gemini-2.5-Flash	0.0506	0.07	0.0194	$p < 0.001$	0.0725	0.1006	0.0281	$p < 0.001$	0.0312	0.0503	0.0191	$p < 0.001$

Table 7: The lexical overlap analysis between questions (original and rewritten) and their golden answers. "Orig" stands for lexical overlap between original questions and golden answers; "Rewr" stands for the lexical overlap between rewritten questions and golden answers; Δ represents the numeric gap between two measurements; "Sig" is the statistical significance obtained via the Wilcoxon signed-rank test.

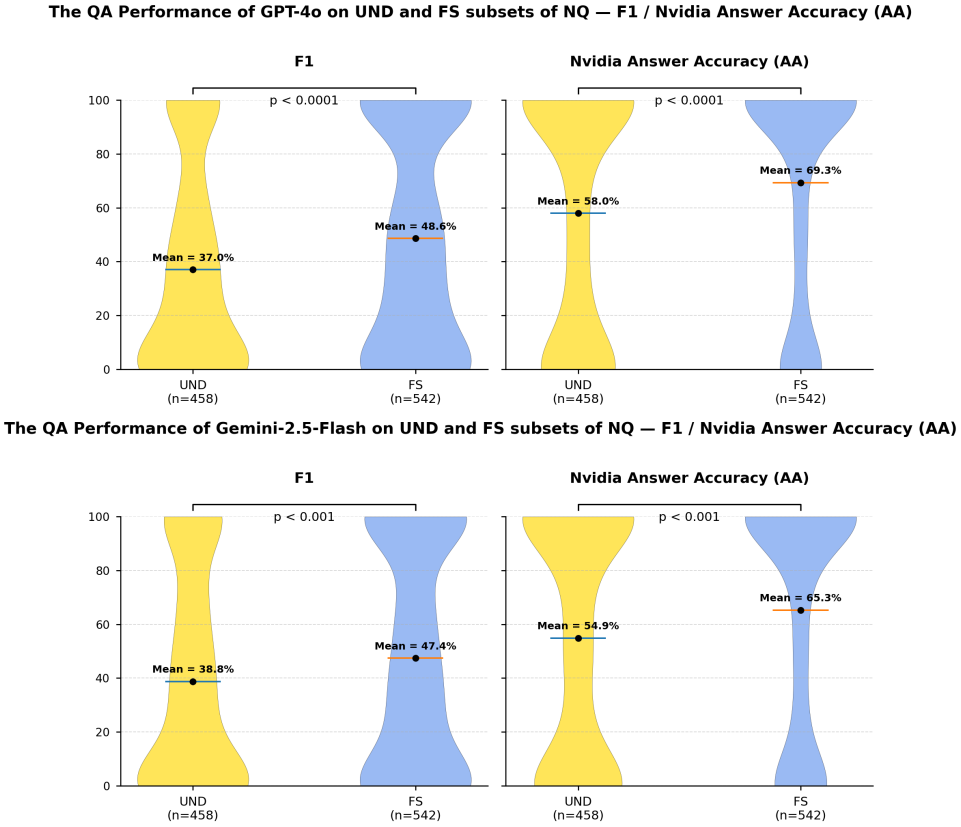

Figure 3: (Step 2) The FS/UND classification results of the **NQ** sample in **QA-ensemble**.

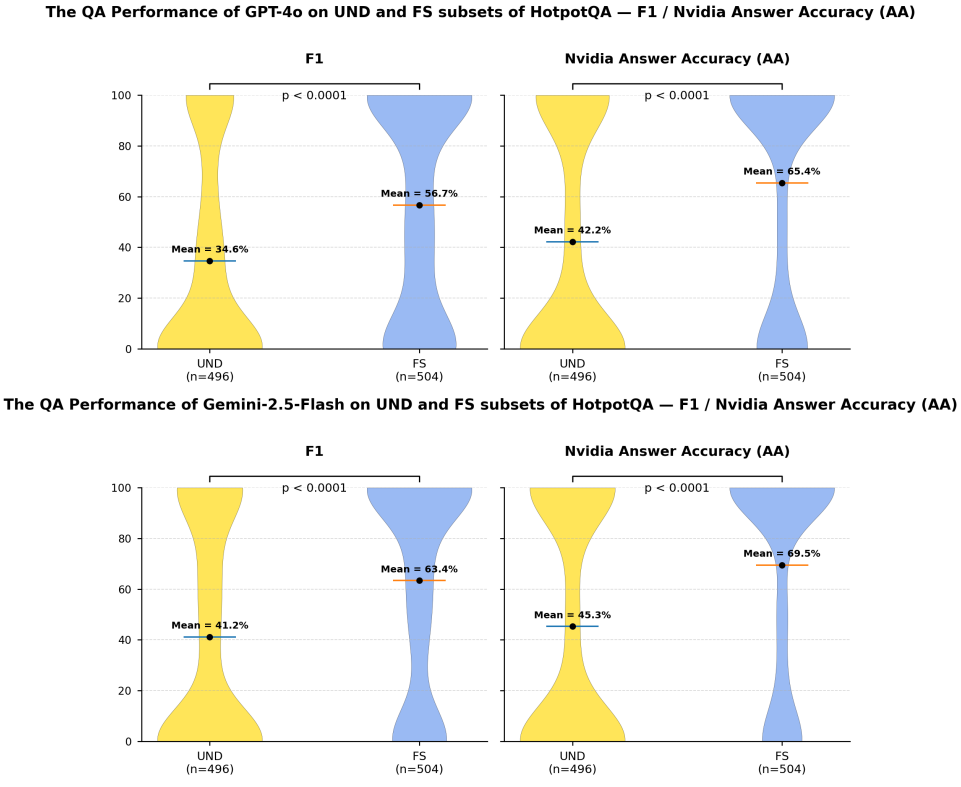

Figure 4: (Step 2) The FS/UND classification results of the **HotpotQA** sample in **QA-ensemble**.

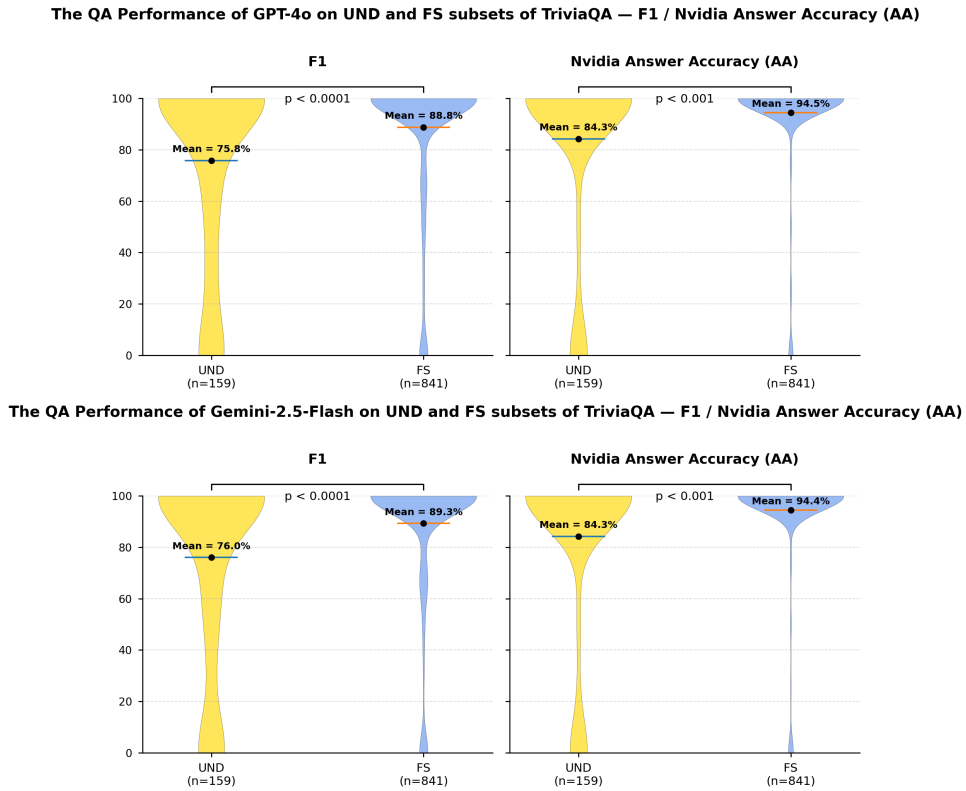

Figure 5: (Step 2) The FS/UND classification results of the **TriviaQA** sample in **QA-ensemble**.

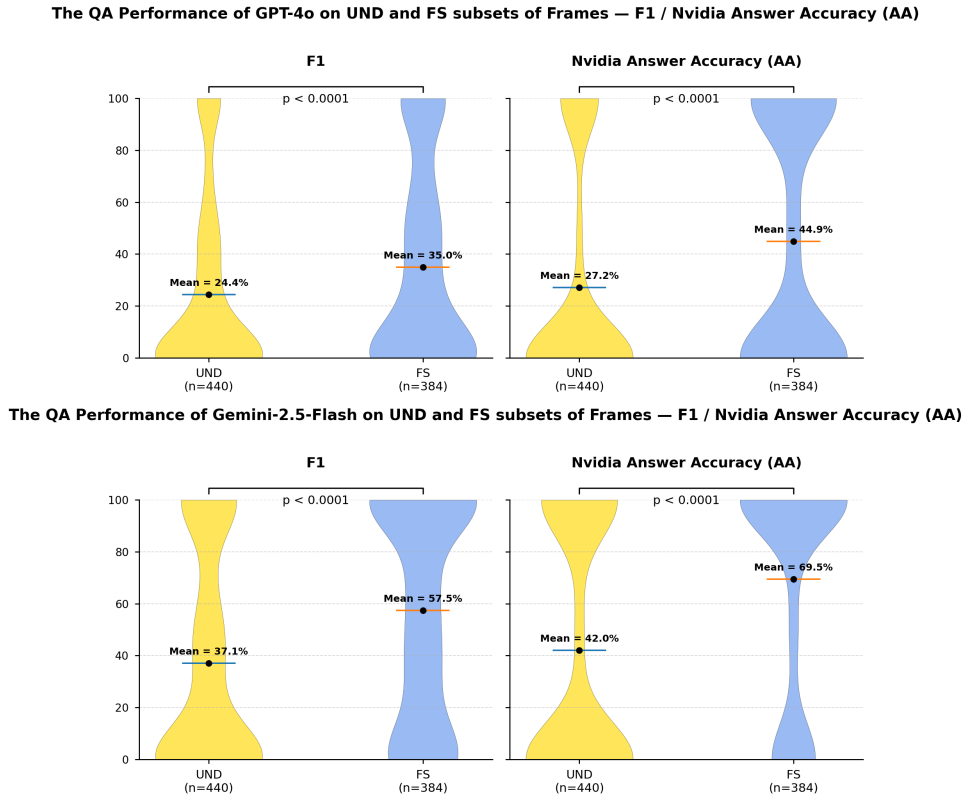

Figure 6: (Step 2) The FS/UND classification results of the **FRAMES** sample in **QA-ensemble**.

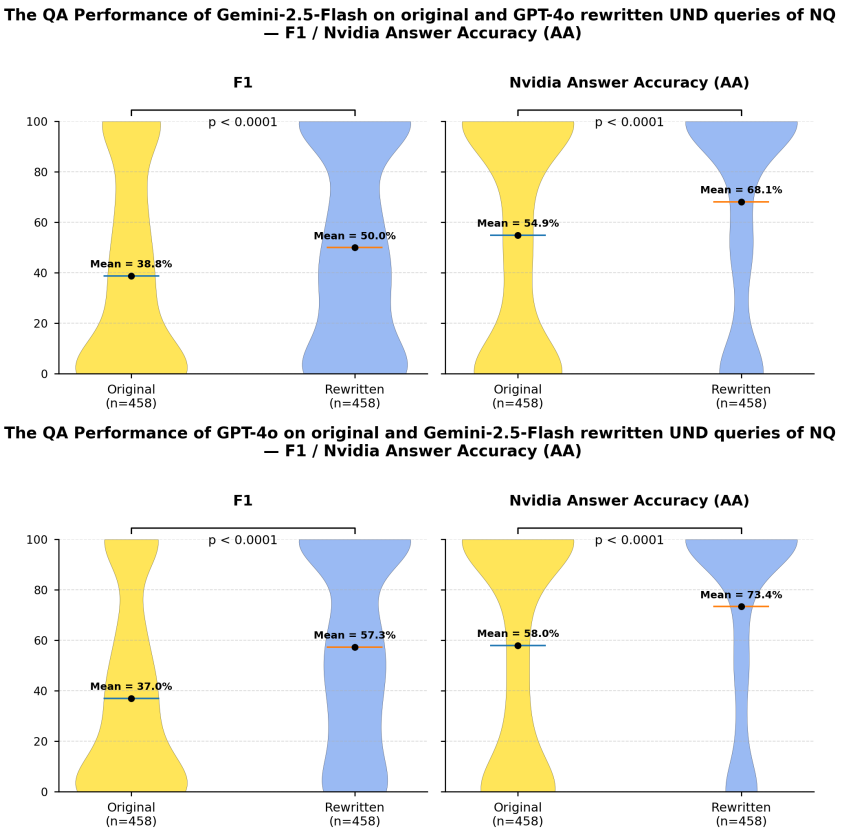

Figure 7: (Step 4) Comparing classification results of the original UND queries and their rewritten counterparts from the **NQ** sample in **QA-ensemble**.

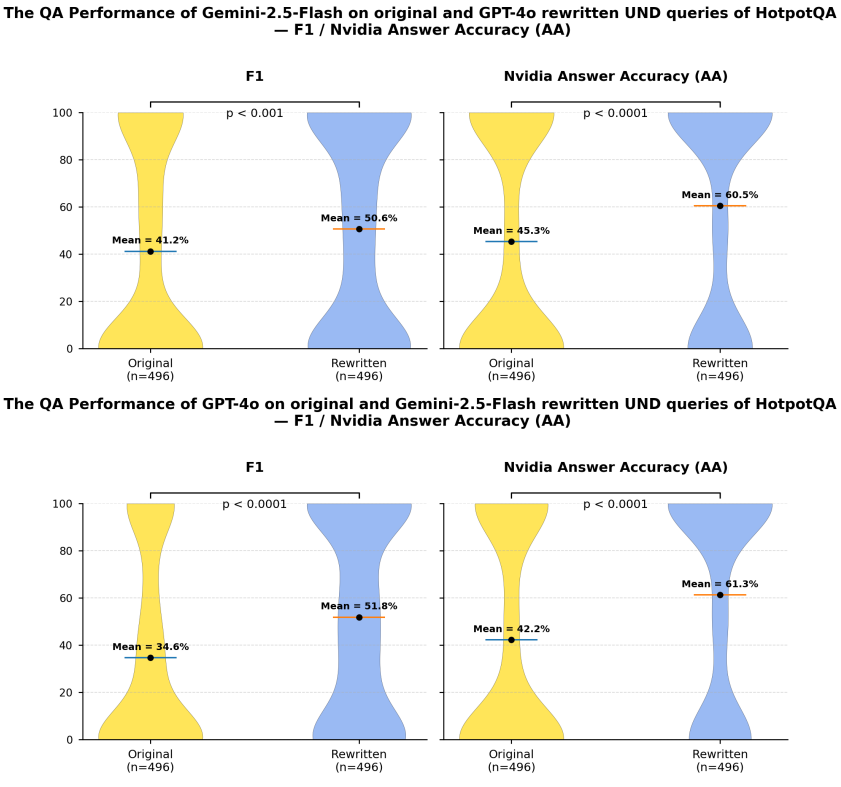

Figure 8: (Step 4) Comparing classification results of the original UND queries and their rewritten counterparts from the **HotpotQA** sample in **QA-ensemble**.

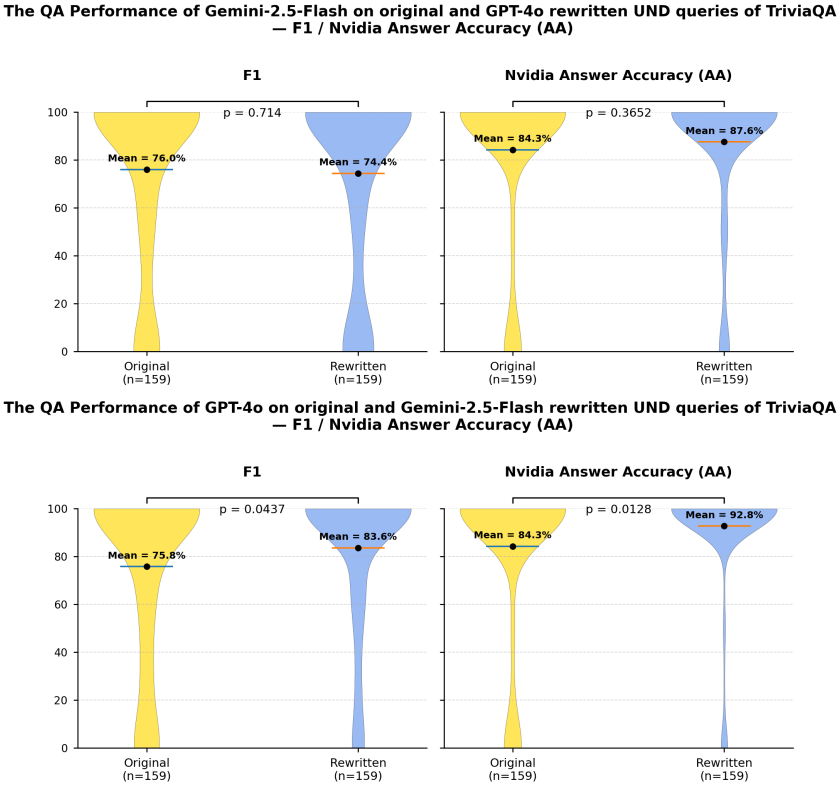

Figure 9: (Step 4) Comparing classification results of the original UND queries and their rewritten counterparts from the **TriviaQA** sample in **QA-ensemble**.

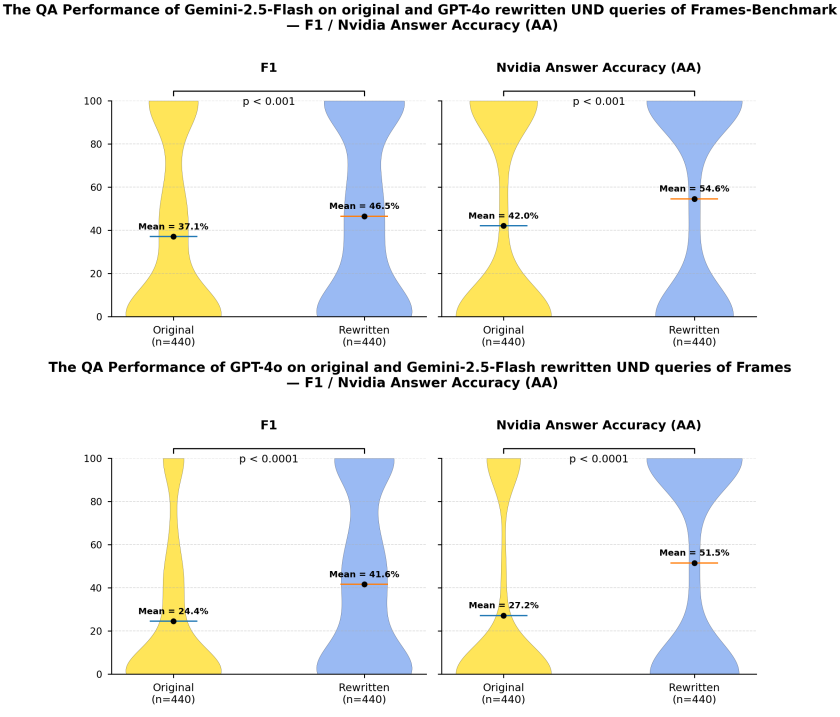

Figure 10: (Step 4) Comparing classification results of the original UND queries and their rewritten counterparts from the **FRAMES** sample in **QA-ensemble**.

Question	Source	QA Model	Golden Answers	Model Answer	Qwen3-4B Pred	Qwen3-4B reasoning
Who is the richest club in the championship?	NQ	GPT-4o	['Manchester City']	['Leicester City']	UND	The query is ambiguous because 'richest' is not defined (e.g., revenue, profit, asset value). The term 'Championship' could refer to different leagues globally (e.g., English FA Cup, Spanish LaLiga, etc.)... Furthermore, the definition of 'richest' varies by metric (broadcast income vs. operating profit vs. shareholder equity)...
Who owns the majority of the US debt?	NQ	Gemini-2.5-Flash	['Japan and China']	['* The U.S. government itself (through intragovernmental holdings)']	UND	...The term 'majority' introduces ambiguity. Key considerations include: 1) Whether 'majority' refers to the largest single holder (e.g., Foreign Governments, Federal Reserve, or Domestic Institutions) or aggregate holdings; 2) Timeframe (current vs historical); 3) Debt instrument specificity (Treasury Bonds, Notes, Bills)...
What was the code name of the landing barge primarily used to provide hot meals to the landing crew?	HotpotQA	GPT-4o	['Operation Neptune']	['HMS Menestheus']	UND	While 'landing barge' is a technical term for a type of watercraft used in amphibious operations, the phrase 'The landing barge' implies a singular, named entity. However, the query does not clarify whether this refers to a historically documented vessel, a fictional construct, or a specific operation...
What colloquial umbrella do Rutgers and the many Ivy league schools fall under?	HotpotQA	Gemini-2.5-Flash	['colonial colleges']	['* Elite schools\n* Top-tier universities\n* Ivy-level schools']	UND	The phrase 'colloquial umbrella' suggests seeking a shared informal label, but Rutgers and Ivy League schools belong to distinct categorizations (public vs. private, regional affiliations). The contradiction between Rutgers (non-Ivy) and Ivy League schools creates ambiguity about what unifying term the user seeks.
Which month of the year came second in the Roman calendar?	TriviaQA	GPT-4o	['april month' 'april' 'mali traven' ...]	['February']	UND	...The query asks for the second month in the 'Roman calendar,' but does not clarify which specific version of the Roman calendar being referenced. The ancient Roman calendar initially began with March as the first month, making February the second. Later reforms, such as the Julian calendar introduced by Julius Caesar, reorganized the calendar to start with January as the first month, making February the second again...
Who wrote the piece of music recognised in much of the Western world as ""The Wedding March"" (or as ""Here Comes The Bride"")?	TriviaQA	Gemini-2.5-Flash	['felix mendelssohn' 'mendelsson bartholdy'...]	['* Richard Wagner']	UND	The query asks about the author of a piece known as 'The Wedding March' or 'Here Comes The Bride.' While these terms are widely recognized, they refer to multiple compositions across different eras and genres. For example: 1) 'The Wedding March' from Wagner's *Lohengrin*, 2) 'Here Comes the Bride' from the 1935 film *The Wizard of Oz* (based on a 1740 French melody), and 3) 'Wedding March' from Mendelssohn's *A Midsummer Night's Dream*...
Who was the King of Siam during the 6th deadliest single-day terrorist attack in U.S. history?	FRAMES	GPT-4o	[King Prajadhipok]	['There was no King of Siam during the 6th deadliest single-day terrorist attack in U.S. history, as Siam was renamed Thailand in 1939.']	UND	... The reference to 'the 6th deadliest single-day terrorist attack in U.S. history' is ambiguous. First, the ranking of terrorist attacks by lethality is subjective and subject to change based on data sources (e.g., differing definitions of 'terrorist attack'). Second, even if the attack's date were determined, Siam's monarchy changed over time—Siam became Thailand in 1939—and the king would vary depending on the year of the attack.
Who developed the first effective vaccine against the disease that killed the father of a famous Hungarian composer born in 1811?	FRAMES	Gemini-2.5-Flash	[Almroth Edward Wright]	['Edward Jenner']	UND	..While the composer and his father's identity are relatively clear (Franz Liszt and his father, Franz Liszt Sr.), the specific disease causing his death remains ambiguous. Historical accounts suggest Franz Liszt Sr. died of typhoid fever, but this is contested. Even if correct, the 'first effective vaccine' for typhoid predates modern times, complicating the answer. ..

Table 8: Additional selected examples of UND queries with incorrect answers by *QA models*

Config	Orig Q	Golden A	Rewr Q	Orig A	Rewr A
From NQ, GPT-4o Rewr, Gemini-2.5-Flash QA	Who owns the majority of the US debt?	['Japan and China']	Which foreign countries currently hold the largest portions of U.S. federal debt, specifically focusing on aggregate holdings, and who are the top two holders?	['* The U.S. government itself (through intragovernmental holdings)']	['Japan, China']
From NQ, Gemini-2.5-Flash Rewr, GPT-4o QA	Who is the richest club in the championship?	['Manchester City']	Based on the most recent financial reports, which club in the English Premier League generated the highest revenue?	['Leicester City']	['Manchester City']
From HotpotQA, GPT-4o Rewr, Gemini-2.5-Flash QA	What colloquial umbrella do Rutgers and the many Ivy league schools fall under?	['colonial colleges']	What historical category do Rutgers University and the Ivy League schools collectively belong to, based on their founding during the colonial period?	[Elite schools / Top-tier universities / Ivy-level schools']	['Colonial colleges']
From HotpotQA, Gemini-2.5-Flash Rewr, GPT-4o QA	What was the code name of The landing barge primarily used to provide hot meals to the landing crew?	['Operation Neptune']	What was the code name for the naval assault phase of the Normandy landings, which encompassed the deployment of numerous landing barges and extensive logistical efforts, including the provision of hot meals to the landing forces?	['HMS Menestheus']	['Operation Neptune']
From TriviaQA, GPT-4o Rewr, Gemini-2.5-Flash QA	Who wrote the piece of music recognised in much of the Western world as "The Wedding March""(or as "Here Comes The Bride")?	['felix mendelssohn'/ 'mendelsson bartholdy'/...]	Who composed the 'Wedding March' from *A Midsummer Night's Dream*, which is widely recognized in much of the Western world as 'The Wedding March'?	['* Richard Wagner']	['Felix Mendelssohn']
From TriviaQA, Gemini-2.5-Flash Rewr, GPT-4o QA	Which month of the year came second in the Roman calendar?	["april month' / 'april' / ...]	In the early Roman calendar, when March was considered the first month of the year, which month came second?	['February']	['April']
From FRAMES, GPT-4o Rewr, Gemini-2.5-Flash QA	Who developed the first effective vaccine against the disease that killed the father of a famous Hungarian composer born in 1811?	Almroth Edward Wright	Who developed the first effective vaccine against typhoid fever, the disease historically believed to have caused the death of Franz Liszt's father, Franz Liszt Sr.?	['Rudolf Weigl']	['Almroth Wright']
From FRAMES, Gemini-2.5-Flash Rewr, GPT-4o QA	Who was the King of Siam during the 6th deadliest single-day terrorist attack in U.S. history?	King Prajadhipok	Who was the King of Siam during the Bath School bombing on May 18, 1927?	['There was no King of Siam during the 6th deadliest single-day terrorist attack in U.S. history, as Siam was renamed Thailand in 1939.']	['King Prajadhipok (Rama VII)']

Table 9: Selected examples of UND queries rewritten by *Rewriter* models, which were initially answered incorrectly by *QA models* but answered correctly in a subsequent round after rewriting.