Studying Rhetorically Ambiguous Questions

Oghenevovwe Ikumariegbe, Eduardo Blanco, Ellen Riloff
Department of Computer Science, University of Arizona
{oaikumariegbe,eduardoblanco,riloff}@arizona.edu

Abstract

Distinguishing between rhetorical questions and informational questions is a challenging task, as many rhetorical questions have similar surface forms to informational questions. Existing datasets, however, do not contain many questions that can be rhetorical or informational in different contexts. We introduce Studying Rhetorically Ambiguous Questions (SRAQ), a new dataset explicitly constructed to support the study of such rhetorical ambiguity. The questions in SRAQ can be interpreted as either rhetorical or informational depending on the context. We evaluate the performance of state-of-the-art language models on this dataset and find that they struggle to recognize many rhetorical questions.

1 Introduction

Question answering (QA) systems and conversational agents, such as chatbots, must handle a wide variety of user questions across diverse settings. Among these are rhetorical questions (RQs)—questions that do not seek an actual answer (Frank, 1990), but instead serve pragmatic functions such as expressing emotion or emphasizing a point (Špago, 2016; Ilie, 1994; Roberts and Kreuz, 1994). In contrast, informational questions (IQs) are genuine requests for information that seek an answer. Distinguishing RQs from IQs is essential, as they elicit fundamentally different responses.

However, this distinction is challenging because RQs and IQs often share similar surface forms. Špago (2016) estimates that only about 15% of rhetorical questions are easily distinguishable from informational questions——a subset we refer to as stylized rhetorical questions. These include questions like *Do pigs fly?* or *How fun is that?*, where the rhetorical nature is obvious due to world knowledge or social conventions.

More commonly, questions are rhetorically ambiguous, meaning the questions are interpreted as informational or rhetorical depending on their context. Consider the question *Why not?* in the following dialogues:

(i) A: Would you like to come along for coffee?
B: Why not?

(ii) A: Don't take the usual route.
B: Why not?

In (i), *Why not* is rhetorical, functioning as an affirmation, similar to saying *Sure*, whereas in (ii), it is informational, seeking to know the reason.

Prior work has recognized the need to distinguish between RQs and IQs (Bhattasali et al., 2015; Kikteva et al., 2024), however, existing datasets contain relatively few rhetorically ambiguous questions. To address this gap, we introduce SRAQ, a dataset containing questions paired with multiple contexts that lead to different interpretations (RQ or IQ), and so more accurately reflects the complexity of real-world language usage.

SRAQ enables a more realistic evaluation, requiring systems to navigate rhetorical ambiguity. Our main contributions are in (a) creating a dataset of rhetorically ambiguous questions in genuine dialogues and (b) evaluating several Large Language Models (LLMs) on it. We find that there is still significant room for improvement on this task.

2 Related Work

Rhetorical questions have been studied extensively in linguistics. Prior research has distinguished them from informational questions and other questions (Frank, 1990; Athanasiadou, 2022), explored their pragmatic functions (Roberts and Kreuz, 1994; Ilie, 1994; Špago, 2020), or examined structural patterns found in some rhetorical questions (Schmidt-Radefeldt, 1977; Schaffer, 2005; Han, 2002; Sadock, 1974; Špago, 2016). However, these works are primarily theoretical and are sourced from movies, plays or magazines. In contrast, our dataset is based on user-authored content.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 30519–30529
November 4-9, 2025

	SWDA-Q[a1]	QT30[b]	Sarcasm-RQ[c]	ERCRQ[d]	IRQSM[e]	SRAQ
Source	SWDA	Talk show	IAC	Twitter	Twitter	Reddit
Domain	Dialogue	Political debates	Political debates	Social media	Social media	Social media
Modality	Spoken	Spoken	Written	Written	Written	Written
Annotations	Human	Human	Heuristics	Human	Heuristics	Human
Context	Prior utterance	Prior utterance	None	Prior tweet	User status	Three posts
Classes	rhetorical, non-rhetorical	rhetorical, pure, assertive	rhetorical, factual	rhetorical, informational	rhetorical, non-rhetorical	rhetorical, informational
# instances	9,411	2,867	2,040	4,997	40,146	971
% RQs	6	14	50	47	40	63

[a] (Bhattasali et al., 2015) [b] (Kikteva et al., 2024) [c] (Oraby et al., 2017) [d] (Zhuang and Riloff, 2020) [e] (Ranganath et al., 2021)

Table 1: Comparison of SRAQ with similar datasets. SRAQ is sourced from Reddit and has longer contexts.

	SWDA-Q[a]	QT30[b]	SRAQ
# instances	9,411	2,867	971
# of rhetorically ambiguous questions	161	16	488
Avg. length of preceding context	14.67	4.52	404.22

[a] (Bhattasali et al., 2015) [b] (Kikteva et al., 2024)

Table 2: Comparison of publicly available datasets for studying rhetorical questions.

On the computational side, Bhattasali et al. (2015) automatically detected rhetorical questions. They worked with the Switchboard Dialog Acts Corpus (Godfrey et al., 1992; Jurafsky et al., 1997) and targeted the subset of dialogue acts pertaining to questions. Zhuang and Riloff (2020) and Ranganath et al. (2021) worked with Twitter data and considered the prior tweet and the user's most recent status message as preceding context respectively. Oraby et al. (2017) distinguished rhetorical questions from factual questions using the Internet Argument Corpus (Abbott et al., 2016), while Kikteva et al. (2024) sourced their data from QT30 (Hautli-Janisz et al., 2022), an annotated corpus of a UK talk show.

Table 1 highlights key differences between our work and prior efforts. Notably, Oraby et al. (2017) and Ranganath et al. (2021) rely on heuristic annotations and do not consider the previous conversation as context. We are the first, to the best of our knowledge, to use Reddit's rich context and take a data-driven approach to obtain rhetorically ambiguous questions.

Table 2 provides a more detailed comparison between SRAQ and previous corpora which were publicly available at the time of writing. Within SRAQ, 50% of instances feature rhetorically ambiguous questions. Notably, when compared to SWDA-Q, the corpus with the highest number of instances, SRAQ contains 3 times as many instances with rhetorically ambiguous questions, despite being only one-tenth of its size. This highlights SRAQ's unique contribution in showing that disambiguation relies on the surrounding context. Furthermore, SRAQ features significantly longer contexts.

3 SRAQ: A Corpus of Rhetorically Ambiguous Questions

In this section, we discuss the creation of SRAQ.

3.1 Retrieving Rhetorically Ambiguous Questions

Our starting point was the `reddit-corpus-small` dataset from Convokit (Chang et al., 2020), which contains discussions from highly active subreddits with high engagement levels. We complement it with other active subreddits likely to contain rhetorically ambiguous questions. This results in a collection of over 24 million posts; see Appendix A for the list of subreddits.

Then, we identify questions in these subreddits using spaCy's `en_core_web_sm` model (Honnibal et al., 2020). We consider any sentence ending with a question mark ("?") a question.

The final step was to identify rhetorically ambiguous questions. We conducted a pilot study to identify the most frequent rhetorically ambiguous questions. Specifically, we annotated a random sample of 50 instances of the 20 most frequent questions in our Reddit corpus. These annotations (Appendix B) revealed that frequent questions have substantially different ratios of rhetorical and infor-

[1] Although we followed the approach detailed in the paper, the number of retrieved instances differs from the 8,515 reported. We keep test set size and label distribution consistent with the original, however.

	Why		What		Why not		So		How	
	RQ	IQ	RQ	IQ	RQ	IQ	RQ	IQ	RQ	IQ
Ratio	0.59	0.41	0.89	0.11	0.70	0.30	0.65	0.35	0.45	0.55
Turn-level features										
a. Average length in tokens	282.69	126.43	215.16	93.27	272.44	113.62	220.60	180.20	258.57	148.89
b. Average length in sentences	15.26	7.89	12.22	6.45	15.19	7.52	12.88	9.51	14.23	9.47
c. Ratio with preceding turn	0.98	0.89	1.00	0.73	0.97	0.97	0.98	0.77	0.98	0.89
d. Ratio with following turn	0.61	0.68	0.67	0.73	0.69	0.86	0.69	0.77	0.61	0.75
Question-level features										
e. Ratio preceded by some other question	0.13	0.19	0.08	0.09	0.21	0.48	0.11	0.26	0.11	0.19
f. Ratio followed by some other question	0.15	0.22	0.23	0.27	0.16	0.21	0.32	0.17	0.30	0.32
g. Ratio that ends the turn	0.09	0.57	0.09	0.36	0.09	0.45	0.06	0.37	0.05	0.26

Table 3: Analysis of the test split in SRAQ. 'Turn-level features' refer to features for the full question turn, while 'Question-level features' refer to those for the rhetorically ambiguous question.

	Rhetorical question
A	Reddit, I really need your help. [...] So that's it. I'm in a clinical major and when I graduated I planned to do a PhD Epi or MPH. **Why?** Because that was the way to help the most amount of people in the short duration of my life.
	Informational question
A	[. . .] I feel that this is literally the antithesis of what this sub is all about.
B	**Care to explain why?** She has a view that it is hard to lose weight and get a healthy routine and she wants us to help her change her view.

Table 4: Examples from the "Why" question ending in SRAQ. We include the preceding turn only if useful to understand the turn containing the question.

mational use. We decided to target five question endings (i.e., last word in a question) likely to indicate rhetorically ambiguous questions: "Why", "What", "Why not", "So", and "How." The final set consisted of 200 anonymized samples from each of these question endings.

3.2 Annotation Process

Annotations were carried out by two graduate students—one is majoring in Computer Science (research focus: NLP) and the other in Linguistics. Before annotating, both annotators completed practice sets to ensure good understanding of the task. They labeled each question, in its context, as either *Rhetorical* or *Informational*, via a custom web interface (see Appendix C). Of the 1,000 instances, 500 were annotated by both people and 500 were annotated by one person.

The doubly annotated set achieved a Cohen's kappa of 0.66, indicating substantial inter-annotator agreement (Artstein and Poesio, 2008). Disagreements were resolved through adjudication. Several disagreements hinged on the subjectivity of rhetorically ambiguous questions in cases where it is rhetorical but one could plausibly interpret it as informational. A representative example is shown in Appendix D.

3.3 Resulting Dataset

We use the doubly annotated portion as a test set, and the singly annotated portion for training and validation. We excluded instances in which the author of the question turn is not actually posing a question, for example, because they are quoting a previous comment. The final dataset contained 384 training, 103 validation and 484 test instances.[2] We present examples from our dataset in Table 4.

Analysis. Table 3 summarizes key statistics of our test set. All question endings have rhetorical and informational labels in different contexts. Rhetorical questions tend to have longer question turns but are less likely to have following turns compared to informational questions (rows (a), (b) and (d)). For the question ending "Why not", informational questions are more often preceded by another question, whereas for "So", rhetorical questions are more frequently followed by a question (rows (e) and (f)). The strongest statistical cue of a question being informational is whether it is the final sentence in the turn (row (g)). However, many informational questions do not end their conversation turn (i.e., the author of the informational question continues the dialogue).

[2]The dataset is available at https://github.com/Abby-OGV/sraq

4 Experiments

We run a series of experiments to evaluate the performance of LLMs on SRAQ. Similar to annotation, the experiments are in the form of a binary classification task to determine the label of a highlighted question—rhetorical or informational. Prior work has suggested that context aids in distinguishing between rhetorical and informational questions. We use a trimmed version of the question turn ("Paragraph"), which is the paragraph containing the question, and thus can include sentences before and after the question.

4.1 Baselines

We establish five baselines: a random baseline, a majority class baseline, a sentiment-based baseline and two baselines which are `RoBERTa`-large models fine-tuned on previous existing datasets.

The sentiment-based baseline is a `RoBERTa`-large model (Liu et al., 2019) trained on TweetEval (Barbieri et al., 2020; Rosenthal et al., 2017) to detect sentiment, with non-neutral sentiments mapping to the rhetorical label and neutral sentiments mapping to the informational label. The QT30 baseline is trained only on instances from the QT30 dataset which are labeled as *pure* (informational) or *rhetorical*, matching our classes. The test set is likewise reduced to instances of those classes.

We detail the hyperparameters used in fine-tuning the baselines in Appendix F.2.

4.2 Prompting LLMs

We conducted prompting experiments on open-source (Grattafiori et al., 2024) and closed-source (OpenAI, 2023; DeepSeek-AI, 2025) instruction-tuned models using the prompt template detailed in Appendix E.

We also ran 2-shot chain-of-thought experiments on the general-purpose LLMs, providing one example with explanation per label. However, performance dropped, likely due to the larger input and the nature of the task. For example, `GPT-4o` averaged 0.65 F1 across 5 runs.

4.3 Fine-tuning

We fine-tune only the general-purpose LLMs, and not the reasoning models, primarily due to cost constraints and the competitive performance of `GPT-4o` to the reasoning models. The details of the fine-tuning are provided in Appendix F.3.

Model	RQ			IQ			Avg.
	P	R	F1	P	R	F1	F1
Baselines							
Majority	0.66	1.00	0.79	0.00	0.00	0.00	0.40
Random	0.66	0.52	0.58	0.34	0.47	0.39	0.49
RoBERTa Sentiment	0.73	0.56	0.64	0.42	0.61	0.50	0.57
RoBERTa SWDA-Q	0.82	0.46	0.59	0.44	0.80	0.56	0.58
RoBERTa QT30	0.81	0.50	0.62	0.44	0.78	0.57	**0.59**
Prompting: General-purpose LLMs							
LlaMa-3.1-8B-Instruct	0.70	0.93	0.80	0.67	0.22	0.33	0.56
GPT-4.1	0.87	0.55	0.68	0.49	0.84	0.62	0.65
GPT-4o-mini	0.76	0.79	0.78	0.57	0.53	0.55	0.66
GPT-4o	0.84	0.76	0.80	0.60	0.72	0.66	**0.73**
Prompting: Reasoning LLMs							
o4-mini	0.87	0.46	0.60	0.45	0.87	0.60	0.60
Deepseek-R1:70B	0.73	0.87	0.79	0.59	0.38	0.46	0.63
o1	0.90	0.65	0.76	0.56	0.87	0.68	0.72
o3	0.87	0.73	0.79	0.60	0.78	0.68	**0.74**
Finetuning on SRAQ							
LlaMa-3.1-8B-Instruct	0.72	0.45	0.55	0.38	0.54	0.44	0.50
GPT-4.1	0.86	0.81	0.84	0.67	0.75	0.71	0.77
GPT-4o-mini	0.91	0.75	0.82	0.64	0.86	0.74	0.78
GPT-4o	0.91	0.77	0.84	0.66	0.86	0.75	**0.79**

Table 5: Classification results. We highlight the best result for each group. Avg. F1 refers to the macro average F1.

5 Main results

Table 5 shows our experimental results in terms of precision (P), recall (R) and macro average F1 scores. We observe a common trend across models: *rhetorical questions are more frequently misclassified as informational than vice versa.*We report misclassification errors across three context settings in Appendix G.

Both the QT30 and SWDA-Q baselines performed worse on SRAQ compared to their original test sets, underscoring SRAQ's emphasis on rhetorical ambiguity and need for stronger contextual understanding. Specifically, QT30 achieved 0.67 F1 on its own test set, but only 0.59 on SRAQ. Similarly, SWDA-Q dropped from 0.79 F1 to 0.58 when evaluated on SRAQ.

Prompting outperforms all baselines, particularly `GPT-4o` and `o3`. However, fine-tuning outperforms prompting, with the exception of `LlaMa-3.1-8B-Instruct`, indicating the benefit of training on SRAQ. Furthermore, fine-tuning narrows the gap between `GPT-4o-mini` and `GPT-4o`, indicating that even smaller models benefit from the task-specific fine-tuning.

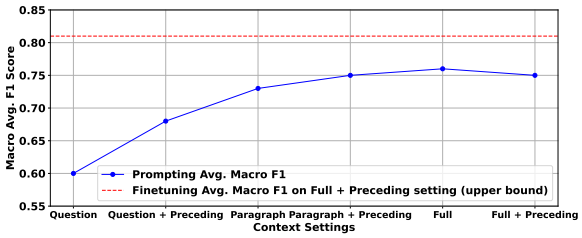

Figure 1: Macro F1 increases with richer context, highlighting the need of context in disambiguation. Question refers to the question only, Paragraph to the paragraph containing the question, and Full to the full question turn. '+ Preceding' includes the preceding context.

	Why	What	Why not	So	How
Fine-tuned GPT-4o	0.80	0.71	0.70	0.84	0.79

Table 6: Macro F1 scores of the fine-tuned GPT-4o on the different question endings. The model performs best on "So" and worst on "Why not".

6 Analysis

We provide a detailed analysis of our experimental results in the following subsections.

6.1 Effect of Context

Figure 1 shows the performance obtained by prompting `GPT-4o` while varying the amount of context provided. The macro F1 improves with more context, except for Full (the full question turn) + Preceding, where it decreases slightly—likely due to the much larger context.

6.2 Performance per Question Endings

Table 6 presents the F1 scores of the best-performing model (fine-tuned `GPT-4o`, Table 5) by question ending. The model performs best on "So" (0.84) and worst on "Why not" (0.70). This is likely due to the greater rhetorical ambiguity of "Why not," consistent with our pilot study (Table 7).

6.3 Incremental Fine-tuning

We further explore the utility of our training set through incremental fine-tuning of `GPT-4o` on portions of the training data. Figure 2 reports F1 scores under two settings: randomly sampled subsets and incremental subsets that accumulate previous subsets. The incremental setting shows roughly linear gains with diminishing returns, except for a drop at 20–40% and a jump at 80–100%. Random sampling shows greater variability, suggesting some sensitivity to specific instances. Overall, performance improves with more training data, peaking with the full training set.

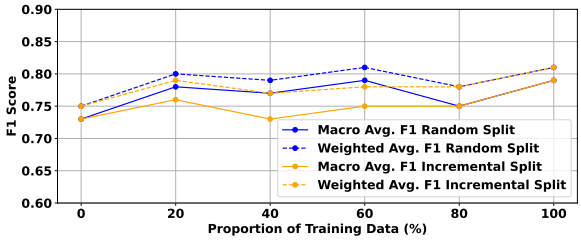

Figure 2: Performance on test set with incrementally fine-tuned `GPT-4o` models.

6.4 Error Analysis on the Fine-tuned GPT-4o

We analyzed the errors made by the fine-tuned `GPT-4o`. The model misclassified 23 IQs and 72 RQs. A manual review of the misclassified questions indicated several patterns:

- **Negative assertions:** RQs expressing challenge, disbelief, or sarcastic undertones were often interpreted as genuine IQs.
- **Shortened context:** Limited context in the Paragraph setting occasionally hindered disambiguation. However, `GPT-4o` trained on longer context settings—Full and Paragraph, + Preceding—show similar performance (0.81 and 0.75 F1 respectively). This suggests that while additional context is beneficial, the model may not always be able to leverage it.
- **Missing cues:** The model occasionally ignored indicative surface cues like repeated punctuations (e.g,. "How?????").
- **Clarification:** Clarification-seeking IQs were sometimes mistaken for RQs, e.g, "Because who says so?".

These findings suggest that while the fine-tuned LLM performs comparatively well, rhetorical ambiguity remains a challenging task for LLMs.

7 Conclusion

We introduce SRAQ, a dataset of rhetorically ambiguous questions from Reddit. Our results show that current state-of-the-art models struggle to recognize many rhetorical questions that have ambiguous question endings. We also run a series of analysis on our experimental results and find that different question endings, as well as, context settings impact performance.

We hope that the SRAQ dataset encourages more work on this challenging problem.

8 Acknowledgements

We thank Matthew Hernandez for his help with the data annotation process.

Several results presented in this work were obtained using computational resources available at the Chameleon testbed, which is supported by the National Science Foundation (Keahey et al., 2020). We also thank the anonymous reviewers for their valuable feedback leading to several additional analysis.

9 Limitations

The primary limitation of this work lies in the dataset size. Given the human effort in constructing SRAQ, we focused on a limited set of question endings to ensure high-quality labels and a manageable scope. Our goal in this paper is not exhaustive coverage, but rather to highlight a gap in prior work regarding rhetorically ambiguous questions. Nonetheless, this limitation can be addressed in future work by expanding the set of question endings and scaling up annotation efforts.

Another limitation of our work is in the use of closed-source models. We prompt and fine-tune closed-source models for several experiments and analysis, incurring costs of $125.

10 Ethics

We are not aware of any ethical violations in our methodology for creating this dataset. However, since the dataset is sourced from Reddit, it may contain offensive language, including slurs or inappropriate terms, as well as implicit or explicit biases. We have taken care to preserve the authenticity of the data for research purposes, but we encourage users of the dataset to be mindful of these aspects when conducting further analysis or model development.

References

Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. 2016. Internet argument corpus 2.0: An SQL schema for dialogic social media and the corpora to go with it. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4445–4452, Portorož, Slovenia. European Language Resources Association (ELRA).

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Angeliki Athanasiadou. 2022. The discourse function of questions. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, page 107–122.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. 2020. TweetEval:Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Proceedings of Findings of EMNLP*.

Shohini Bhattasali, Jeremy Cytryn, Elana Feldman, and Joonsuk Park. 2015. Automatic identification of rhetorical questions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 743–749, Beijing, China. Association for Computational Linguistics.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. ConvoKit: A toolkit for the analysis of conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60, 1st virtual meeting. Association for Computational Linguistics.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jane Frank. 1990. You call that a rhetorical question?: Forms and functions of rhetorical questions in conversation. *Journal of Pragmatics*, 14(5):723–738.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, et al. 2024. The llama 3 herd of models.

Chung-hye Han. 2002. Interpreting interrogatives as rhetorical questions. *Lingua*, 112(3):201–229.

Annette Hautli-Janisz, Zlata Kikteva, Wassiliki Siskou, Kamila Gorska, Ray Becker, and Chris Reed. 2022. QT30: A corpus of argument and conflict in broadcast debate. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages

3291–3300, Marseille, France. European Language Resources Association.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Cornelia Ilie. 1994. *What else can I tell you?: A Pragmatic Study of English Rhetorical Questions as Discursive and Argumentative Acts*. Almqvist & Wiksell International.

Daniel Jurafsky, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema. 1997. Automatic detection of discourse structure for speech recognition and understanding. In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 88–95. IEEE.

Kate Keahey, Jason Anderson, Zhuo Zhen, Pierre Riteau, Paul Ruth, Dan Stanzione, Mert Cevik, Jacob Colleran, Haryadi S. Gunawi, Cody Hammock, Joe Mambretti, Alexander Barnes, François Halbach, Alex Rocha, and Joe Stubbs. 2020. Lessons learned from the chameleon testbed. In *Proceedings of the 2020 USENIX Annual Technical Conference (USENIX ATC '20)*. USENIX Association.

Zlata Kikteva, Alexander Trautsch, Steffen Herbold, and Annette Hautli-Janisz. 2024. Question type prediction in natural debate. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 624–630, Kyoto, Japan. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

OpenAI. 2023. Gpt-4 technical report. ArXiv:2303.08774.

Shereen Oraby, Vrindavan Harrison, Amita Misra, Ellen Riloff, and Marilyn Walker. 2017. Are you serious?: Rhetorical questions and sarcasm in social media dialog. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 310–319, Saarbrücken, Germany. Association for Computational Linguistics.

Suhas Ranganath, Xia Hu, Jiliang Tang, Suhang Wang, and Huan Liu. 2021. Identifying rhetorical questions in social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1):667–670.

Richard M. Roberts and Roger J. Kreuz. 1994. Why do people use figurative language? *Psychological Science*, 5(3):159–163.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518.

Jerrold M Sadock. 1974. *Toward a Linguistic Theory of Speech Acts*. Academic Press.

Deborah Schaffer. 2005. Can rhetorical questions function as retorts?: Is the pope catholic? *Journal of Pragmatics*, 37(4):433–460.

Jürgen Schmidt-Radefeldt. 1977. On so-called 'rhetorical' questions. *Journal of Pragmatics*, 1(4):375–392.

Džemal Špago. 2020. Rhetorical questions as aggressive, friendly or sarcastic/ironical questions with imposed answers. *ExELL*, 8(1):68–82.

Yuan Zhuang and Ellen Riloff. 2020. Exploring the role of context to distinguish rhetorical and information-seeking questions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 306–312, Online. Association for Computational Linguistics.

Džemal Špago. 2016. Rhetorical questions or rhetorical uses of questions? *ExELL*, 4.

A Subreddits Used in Creating the Dataset

The following subreddits were used in creating the dataset:

- The 100 subreddits of Convokit's `reddit-corpus-small` dataset.
- ApplyingToCollege
- europe
- changemyview
- Cornell
- AskHistorians
- legaladvice

B Frequency of the Top 20 Most Common Questions in the Reddit Corpus

We present the top 20 most common questions, sorted in decreasing order of frequency, as well as the ratio of rhetorical questions present in the sample of 50 in Table 7.

We do not take exactly the top 5 questions in our selection since we pay attention to rhetorical ambiguity. For example, although "Really?" appeared among the top five most-frequent question ending, we found that it lacked sufficient rhetorical ambiguity to be included in the dataset.

Similarly, "So what?" can be found under "What?" due to the use of question endings, hence

Question	Count	Ratio of RQs
Why?	37,730	0.52
Any questions or concerns?	24,457	0.00
What?	23,554	0.78
Really?	21,447	0.80
Why not	7,644	0.64
So?	6,755	0.42
Source?	6,322	0.06
So what?	6,174	0.78
How?	6,143	0.28
What should I do?	6,082	0.00
Right?	6,038	0.54
Seriously?	5,494	0.90
How so?	5,454	0.44
What can I do?	5,174	0.02
Huh?	4,904	0.84
Is this legal?	4,806	0.08
No?	4,452	0.71
What do you mean?	4,216	0.21
What are you talking about?	4,057	0.90
What do I do?	3,509	0.02

Table 7: Top 20 most common questions that guided our data-driven approach in selecting the final set of question endings. The count refers to the count in the Reddit corpus, while the ratio is the proportion of rhetorical questions in the samples of 50.

we skipped over it in our choice of question endings.

C Annotation Interface

Annotations were done on a website. The annotation guidelines were provided on the right of the screen and annotations on the left side in sets of 50. The interface was color-coded and separated into blocks to help with distinguishing turns.

The guidelines also contained clear definitions and several examples to ensure the quality of the annotations. 50 instances were presented at each given time to reduce annotation fatigue with 10 instances each for the question endings. We show a sample image of the annotation interface in Figure 3.

D A Representative Example of a Resolved Annotator Disagreement

In this section, we present an example of annotator disagreement that was resolved through adjudication. The question is presented in <ques></ques> markers.

> *Preceding Turn:* I'm not saying there is enough grazeland to satisfy the demand for beef. I am saying that if you don't like factory farming support free range. I don't think the supply of beef should be as high as it is and I think the price of beef should be higher.
>
> *Question Turn:* <ques>If there isn't enough grazeland... then... free range... how?</ques>

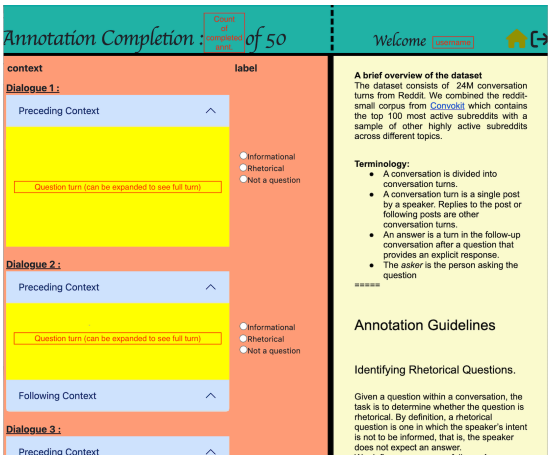

Figure 3: Web interface for annotation.

An initial interpretation would be that the question turn author is interested in actually knowing how the free range farming would be implemented, but a closer look at the previous context would reveal that the author is expressing disbelief because the reasoning does not make sense. In a sense: Without enough grassland, how can there be free range chickens?!

This is why the annotators went through a thorough adjudication afterwards to fully understand each other's interpretation and how the context supports their choice.

E Prompts Used

We experimented with several prompt formulations and selected the best-performing one based on model results. The prompts shown in Table 8 fall into three categories:

- **With definitions:** Determine whether the question is Rhetorical or Informational, with definitions provided (Prompt 0).
- **Without definitions:** Same task as above, but without definitions (Prompt 1).
- **Yes/No classification:** Determining whether the question explicitly expects an answer (Prompt 2).

Additionally, we experimented with two asymmetric prompts ——each focusing on only one of the two labels (i.e. "Determine whether the sentence is a rhetorical question." and vice versa). These

	System prompt	User prompt
0	You are a context-aware expert that has been provided with a dialog snippet with a highlighted sentence. Your task is to determine what category the highlighted sentence belongs to: - "Informational": The sentence is a question that explicitly seeks information or an answer. - "Rhetorical": The sentence is a question that is not intended to elicit an explicit answer but serves a pragmatic purpose. The expected output is a JSON object with the following structure: "label": "Rhetorical" or "Informational". "explanation": A short justification (1–2 sentences, max 30 tokens) for your label choice, considering phrasing, context, and intended effect.	Determine whether the sentence enclosed in the <ques> </ques> markers is "Informational" or "Rhetorical". Answer "Rhetorical" if it is a rhetorical question and "Informational" if it is an information-seeking question. Context: <context>
1	You are a context-aware expert that has been provided with a dialog snippet with a highlighted sentence. Your task is to determine what category the highlighted sentence belongs to: - "Informational": The sentence is an information-seeking question. - "Rhetorical": The sentence is a rhetorical question. The expected output is a JSON object with the following structure: "label": "Rhetorical" or "Informational". "explanation": A short justification (1–2 sentences, max 30 tokens) for your label choice, considering phrasing, context, and intended effect.	Determine whether the sentence enclosed in the <ques> </ques> markers is "Informational" or "Rhetorical". Answer "Rhetorical" if it is a rhetorical question and "Informational" if it is an information-seeking question. Context: <context>
2	You are a context-aware expert that has been provided with a dialog snippet with a highlighted sentence. Your task is to determine what category the highlighted sentence belongs to: - "Yes": The sentence is a question that seeks an explicit answer - "No": The sentence does not seek an answer; instead, it serves a pragmatic purpose. The expected output is a JSON object with the following structure: "label": "Yes" or "No". "explanation": A short justification (1–2 sentences, max 30 tokens) for your label choice, considering phrasing, context, and intended effect.	Determine whether the sentence enclosed in the <ques> </ques> markers is a question that explicitly seeks an answer. Answer "Yes" if the sentence is a question that seeks an answer, and "No" if it does not seek an answer but is for pragmatic purpose. Context: <context>

Table 8: Prompts tested for the classification task. Prompt 0, which includes definitions, produced the best results and was used in all subsequent prompting experiments.

led to over-classification of the label mentioned in the prompt, highlighting a priming bias toward the named category.

F Reproducibility

We took several measures to ensure reproducibility. A fixed random seed (42) was used across all experiments, including dataset splits, model training, model prompting, and hyperparameter tuning.

F.1 Prompting

We disable sampling and set the temperature to zero to minimize variance in model outputs. Combined with the fixed seed, we aimed to ensure fairly deterministic generation across runs.

For Deepseek-R1, we use a distilled version with 70B parameters available on Ollama.[3]

F.2 Fine-tuning on RoBERTa

For hyperparameter search, we used Optuna (Akiba et al., 2019) with the default Tree-structured Parzen Estimator (TPE) sampler to optimize hyperparameters. The optimization aimed to maximize the macro-average F1 of the validation set for the SWDA-Q and QT30 baselines, and minimize the validation loss for the Sentiment baseline. We inserted special markers around the target question —<ques> and </ques>—to clearly delineate the question to be classified for the SWDA-Q and QT30 baselines. The batch size was selected from a fixed set {8, 16, 32}, while other hyperparameters such as learning rate and number of epochs are sampled from defined ranges.

We also use early stopping after 3 epochs of no improvements for the actual fine-tuning. The final

Baseline	Hyperparameters
RoBERTa Sentiment	learning rate: 2.4e-5, batch size: 32, number of epochs: 2
RoBERTa SWDA-Q	learning rate: 1.49e-5, batch size: 16, number of epochs: 10
RoBERTa QT30	learning rate: 2.2e-5, batch size: 16, number of epochs: 9

Table 9: Hyperparameters used for fine-tuning the RoBERTa-large baselines.

[3] https://ollama.com/library/deepseek-r1:70b

hyperparameters are detailed in Table 9.

F.3 Fine-tuning on General-Purpose Models

For GPT models, we perform fine-tuning on our training set using the OpenAI Fine-Tuning API for 3 epochs, with a batch size of 1 and a random seed of 42.

For LlaMa-3.1-8B-Instruct, we apply QLoRA fine-tuning using the HuggingFace Trainer API and PEFT library for 7 epochs, with a batch size of 8, a learning rate of 5e-5 and the same random seed of 42. In both cases, we use the same prompt template as in the prompting experiments, except that we omit the requirement for the model to generate a label explanation.

G Label Error Shift Across Contexts

We analyzed how misclassification errors shift across three context settings: Paragraph, Paragraph + Preceding, and Full. Figure 4 shows the distribution of misclassified informational questions (informational questions predicted as rhetorical) and misclassified rhetorical questions (rhetorical questions predicted as informational). We find that `LlaMa-3.1-8B-Instruct` tends to misclassify more informational questions as rhetorical across all context settings, while `GPT-4o` and `GPT-4o-mini` show a more balanced error distribution when given more context. The reasoning models, on the other hand, tend to misclassify more rhetorical questions as informational.

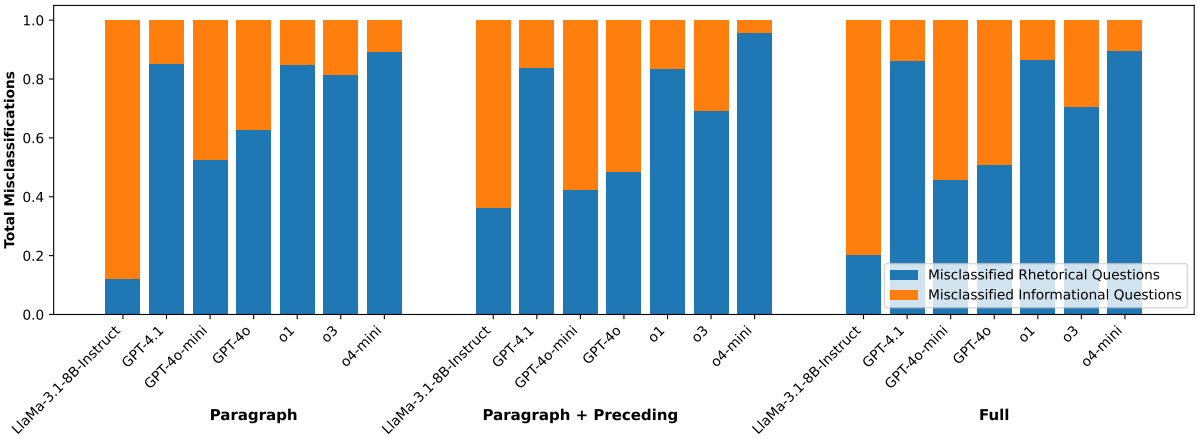

Figure 4: Label error distribution shifts for models across the different context settings.