BLUFF-1000: Measuring Uncertainty Expression in RAG

Anonymous submission

Abstract

Retrieval-augmented generation (RAG) systems often fail to adequately modulate their linguistic certainty when evidence deteriorates. This gap in how models respond to imperfect retrieval is critical for the safety and reliability in real-world RAG systems. To address this gap, we propose BLUFF-1000, a benchmark systematically designed to evaluate how large language models (LLMs) manage linguistic confidence under conflicting evidence to simulate poor retrieval. We created a novel dataset, introduced two novel metrics, and calculated comprehensive metrics to quantify faithfulness, factuality, linguistic uncertainty, and calibration. Finally, we tested generation components of RAG systems with controlled experiments on seven LLMs using the benchmark, measuring their awareness of uncertainty and general performance. Our findings uncover a fundamental misalignment between the linguistic expression of uncertainty and source quality through seven state-of-the-art RAG systems. We recommend that future RAG systems refine uncertainty-aware methods to convey confidence throughout the system transparently.

Introduction

Although RAG implementation has been proven to bring about improvements in answer accuracy [18], concerns have been raised about the lack of robustness and expression of confidence when large language models (LLMs) are supplied with misleading or inaccurate sources [22]. The abundance of misinformation found online [29] further exacerbates this issue.

When encountering poor retrieval environments, LLMs often remain numerically overconfident, maintaining high predicted probabilities despite poor or contradictory evidence [24]. This overconfidence extends to linguistic confidence, as models cannot express uncertainty appropriately [11]. Existing benchmarks measure accuracy and calibration, given high-quality retrieval results [13]. However, they do not evaluate how a model systematically responds when source quality degrades. This flaw in the ways models respond to imperfect retrieval is critical for the safety and reliability of a real-world RAG system.

This dilemma brings us to our two hypotheses:

H1: When supporting retrieved documents are sparse, irrelevant, or contradictory, models will still deliver answers with unwarranted certainty, exhibiting verbal overconfidence not justified by evidence.

H2: Models do not hedge when truly uncertain. Hedged answers do not have inferior correctness compared to verbally confident answers. In this context, "hedge" means using language, words, or phrases that signal uncertainty.

Using BLUFF-1000 and a comprehensive set of metrics, we perform a thorough performance assessment of hypotheses and all models. Our key contributions include the following:

1. The creation of data set BLUFF-1000 to measure a model's ability to express uncertainty in responses. The dataset includes 500 questions, each of which contains two source sets, allowing for 1000 total evaluation instances.
2. The creation of a novel metric to measure verbal uncertainty, named the Verbal Uncertainty Index (VUI). VUI quantifies the frequency of hedge words in a response with respect to accuracy and identifies the extent to which a model uses linguistic uncertainty when necessary.
3. The creation of a novel metric, labeled Ambiguity Sensitivity Index (ASI), to quantify changes in a model's confidence changes when evidence shifts from clear to ambiguous.
4. Linguistic confidence expression and source quality across seven state-of-the-art LLMs were found to be fundamentally misaligned.

Related works

RAG Benchmarks

In previous RAG benchmarks, popular evaluation aspects included factuality, faithfulness, and retrieval quality. CRAG [32] and FRAMES [17] assess the truthfulness of responses and frequency of hallucinations to evaluate the factuality aspect of models. Additionally, retrieval quality is evaluated in GaRAGe [27] and LegalBench-RAG [23]. RAGAS provides an automated framework for the evaluation of RAG's retriever and generator segments to assess faithfulness, answer relevance, and context relevance [8].

Real-world limitations

One major flaw of RAG systems is the inability to handle the imperfect aspects of real-world sources, namely source reliability and the prevalence of counterfactual information.

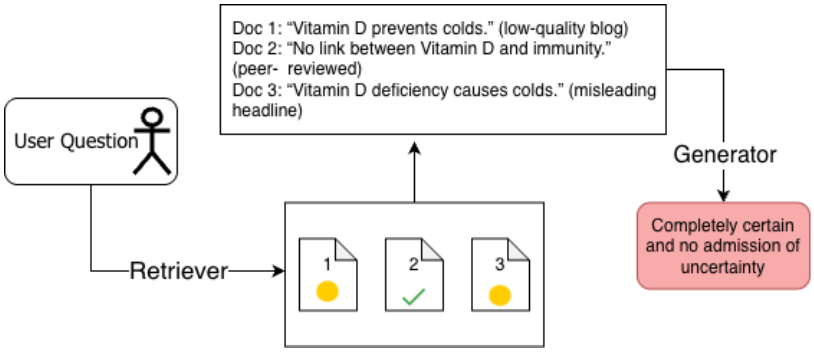

Figure 1: **Example of a RAG pipeline producing an overconfident answer.** The retriever surfaces mixed-quality sources of varying credibility, but the LLM produces an overconfident answer without hedging or uncertainty.

[21, 5, 4]. Solutions have been proposed to tackle this shortcoming. SGIC was established as a framework that improves calibration by calculating separate uncertainty scores after document retrieval and after response generation [6]. The model subsequently utilizes these uncertainty estimates to adjust the confidence of future responses. Similarly, UncertaintyRAG quantifies retrieval chunk uncertainty scores so that models can attach lower confidence to bad-quality retrievals, improving expressions of uncertainty [19]. An outline of established approaches for handling noisy source environments is also provided in section "Handling Noisy Environments" of the Appendix.

Methodology

Dataset generation

As discussed in related literature [28, 2, 26], RAG is cited for its improvements to domain-specific question answering. Thus, our questions span 10 domains and require knowledge of specific fields, which ensures that external sources are essential to answer generation. The Open-AI GPT-4o model is utilized to assist in question generation and source retrieval. We generated a complete dataset of 500 domain-specific, knowledge-intensive questions. This process is outlined in "Question-Answer Pair Generation Outline" and visualized in Figure 9 (see Appendix). Additionally, we collected and classified sources scraped from the internet to add to appropriate source sets. This process is outlined in "Obtaining Sources Outline" and visualized in Figure 10 of the Appendix. To reference the question and source set formats, see Figures 5, 6, & 7 in the Appendix.

Evaluation

See Figure 8 in the Appendix for the evaluation prompt. We chose 7 LLMs (DeepSeek, GPT-3.5, GPT-4o, Llama-3.3, Llama-4, Mistral, Qwen2.5) to evaluate due to diversity in architectures, training paradigms, and strengths [7, 3, 20, 10, 14, 31]. Evaluation aspects included general performance and uncertainty awareness. A model performance comparison can be found in Table 1 of the Appendix.

Metrics

In this benchmark, we derive internal numeric confidence from token-level probabilities using p(true) confidence estimation following prior work in *Language Models (Mostly) Know What They Know*. [15] Additional metric definitions and formulas are found in namesake sections of the appendix.

ASI We propose a novel metric to answer our first hypothesis: **Ambiguity Sensitivity Index (ASI)** which evaluates whether the model appropriately lowers confidence and raises hedging when faced with ambiguous sources.

$$\text{ASI} = \frac{\text{Confidence Sensitivity} + \text{Hedging Sensitivity}}{2}$$

Confidence Sensitivity measures the difference in confidence on clear and ambiguous questions. If the Confidence Sensitivity is negative, a ×2 penalty is applied to emphasize the flawed behavior that is correlated with a negative score, which signifies the model's internal numeric confidence increased as the sources' quality improved. **Hedging Sensitivity** measures the difference in hedging rates between ambiguous and clear sources which optimally should be high and positive. **Hedging Rate** measures the ratio of hedge words/phrases (e.g., perhaps, could be) to total words.

VUI For our second hypothesis, we propose the **Verbal Uncertainty Index (VUI)**. This novel metric evaluates how precisely the model uses hedged language, measuring the F1-score of hedge detection.

$$\text{VUI} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

We calculate this using a confusion matrix approach. **True Positives (TP)** occur when the model uses hedging language and the answers are incorrect. **False Positives (FP)** occur when the model uses hedging language, but the answers are correct. **True Negatives (TN)** take place when the model doesn't use hedging language and the answers are correct. **False Negatives (FN)** are when the model doesn't use hedging language, but the answers are incorrect. From this confusion matrix, hedge recall and hedge precision are calculated.

Results

Uncertainty Awareness

ASI reveals significant variation in models' ability to adjust their confidence (both internal and linguistic) based on source quality. A higher ASI indicates better sensitivity and ideal behavior. Llama-4-Scout achieved the highest ASI of 0.067, demonstrating that when faced with worse sources, this model expressed linguistic uncertainty and lower confidence to a greater, but still not valid extent. Among models, GPT-3.5 Turbo showed the worst performance of -0.131, indicating **strong backwards uncertainty behavior**. Backwards uncertainty behavior signifies an increase in hedging when sources switch from ambiguous clear, which is not ideal. 5 out of 7 other models exhibited negative ASI values, suggesting that overall confidence increased rather than decreased when given ambiguous sources.

This leads us to our diagnostic metric: *Source Set On Hedging*, perhaps the most concerning finding. *Source Set On Hedging* is the population-level version of *hedging sensitivity* which averages differences in hedging rates. 6 out of 7 models hedge less when given the ambiguous source set, which is a direct contradiction of the appropriate uncertainty expression displayed by humans. DeepSeek-V3.1 stands as the sole exception (+0.100), appropriately increasing hedging with ambiguous sources. Alternatively, GPT-3.5 Turbo showed the worst source set on hedging (-1.4); this model hedged significantly less when sources were ambiguous.

Beyond numerical confidences and linguistic hedging behavior, the models also demonstrated varying degrees of faithfulness with Qwen2.5-72B-Instruct-Turbo achieving the highest answer correctness (78.5%) and Llama-4-Scout the highest faithfulness (74.9%). The rest of the models indicated fair faithfulness scores ranging from 0.644, GPT-3.5 Turbo at the bottom, to 0.75 (Llama-4-Scout).

Hedging Behavior

Models revealed a systematic pattern of high hedge recall (0.65-1.00) but low hedge precision (0.02-0.20), indicating that they hedge frequently but inappropriately. DeepSeek-V3.1 was the only model with a positive source set on hedging (+0.100), highlighting increased hedging with ambiguous sources. The best hedging quality was demonstrated by Llama-3.3-70B, which had the greatest VUI (0.327), while Qwen2.5-72B had the lowest VUI (0.046). Interestingly, despite having weak VUI, Qwen2.5-72B had the highest answer correctness (78.5%), indicating that there is no discernible relationship between factual accuracy and hedging behavior. "Why Language Models Hallucinate" highlights that existing benchmarks reinforce hallucination by penalizing uncertainty and refusals. [16]. To combat this, we calculate the refusal sensitivity and missed refusals as well as prompting the model to not guess unless it believes that its confidence is over 0.2 (See Figure 8 in the Appendix). If the model believes that it will answer poorly, it has the option to refuse to answer and will receive 0.4 as the answer accuracy for incentive. Evaluations revealed significant variation in the ability handle refusals. GPT-3.5 Turbo demonstrated the highest refusal sensitivity (0.431). This suggests strong uncertainty awareness, as the model appropriately declined to provide answers when source evidence was insufficient. GPT-4o showed fair refusal sensitivity (0.206) as well, indicating robust uncertainty detection capabilities.

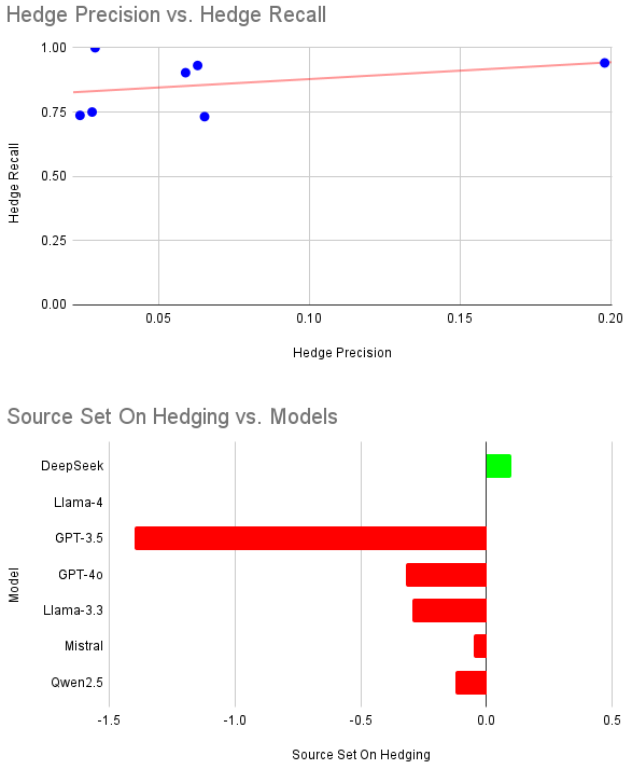

Figure 2: Comparison of hedging-related metrics across models.

Discussion

Evaluations reveal a consistent pattern of **weak uncertainty awareness**. It is evident that LLMs currently fail to express appropriate linguistic signaling when they should output indicators of skepticism. This is highlighted with both a) the hedging rates when source quality degrades (supporting hypothesis 1) and b) the hedging rates relative to answer correctness (supporting hypothesis 2). This is crucial for RAG deployments, as users are easily misled by confident language in wrong answers or a lack of confident language in correct answers.

Understanding when to hedge

Ultimately, these evaluations show the fundamental disconnect between the way generator components in RAG systems express confidence through language and the actual quality of the retrieved evidence. For example, for one question on average, most models expressed more uncertainty in the clear set version of the question than in the ambiguous set. This is demonstrated through negative *Source Set On Hedging* scores. This is counterintuitive, since the question

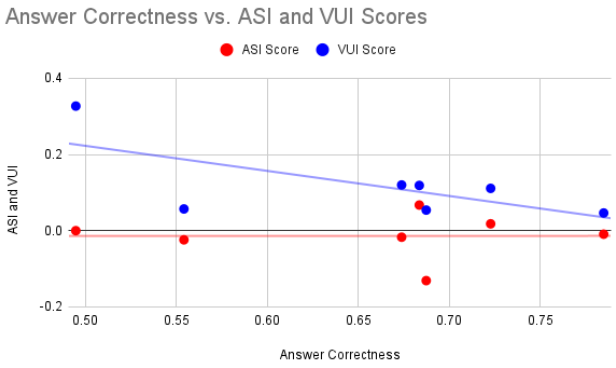

Figure 3: Performance Trade-offs — Accuracy does not correlate with uncertainty awareness (ASI).

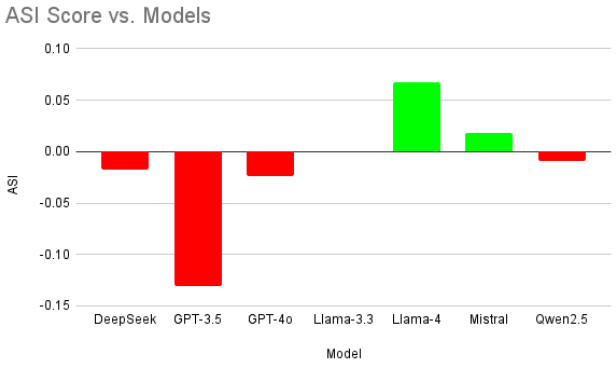

Figure 4: Uncertainty Awareness Comparison — Positive ASI indicates appropriate uncertainty awareness; negative indicates overconfidence.

set with ambiguous, contradictory, and unreliable information is harder to answer and **should** bring about less certain responses.

This counterintuitive linguistic behavior may be attributed to the fact that the model is unaware of the question's difficulty. Hence, the model may not recognize when unreliable or contradictory evidence is retrieved or when the retriever falsifies missing evidence by hallucinating the gaps in evidence. Additionally, the model may recognize the question is difficult, lower its numeric confidence, but still not deliver an appropriate hedging use. This can be seen in the discrepancies between *source set on hedging/hedging sensitivity* and *confidence drop*, indicating that they are not always symmetrical.

Calculated VUI scores further support this and indicate that models struggle to decipher when to hedge and when to remain confident in generation. Although models demonstrate high numerical overconfidence in some responses, they avoid confident language. In other words, models often over-hedge in their confident and accurate answers. In addition to not expressing enough confidence in correct answers, the fact that models did not express any uncertainty when generating poor accuracy responses was another prevalent issue. Furthermore, models that are more correct in their answers are also more likely to use confident language inappropriately. This indicates that training processes that improve factual accuracy inadvertently teach models to express more certainty than necessary.

A prime example of this is in *figure 11*. In this example, Mistral-7B receives a question followed by its coordinating ambiguous source set. By nature, the question is harder to answer, and therefore the generation should have more hedging language signaling uncertainty. The model reports an internal numeric confidence of 0.99 with an accuracy of 0.4 but more importantly, the model response contains zero hedging terms. So not only was the model overconfident numerically when retrieval took a toll, but it was also overconfident linguistically. This is complemented by the fact that confidence did not drop from the corresponding clear set version of this question.

Refusal-friendly evaluation

Current benchmarks penalize "I don't know" responses and treat these types of responses as failures. Over time, these gaps in LLM evaluation created biases toward overconfident linguistic generation. The bias is further compounded when models are incentivized to answer questions over not answering, and therefore give unfaithful responses, creating unsupported claims contributing to hallucination.[16] Our results show what happens with this current style of training: models like Qwen2.5 get high accuracy but are terrible at hedging and admitting doubt, likely because they've been trained to never say 'I don't know' to accommodate for these flawed benchmarks.

The faithfulness scores observed across models (0.644-0.749), as well as occasional poor refusal sensitivity scores (-0.038-0.431), reflect this fundamental issue in prior LLM evaluation. By incorporating faithfulness calculations with refusal incentivization, we combat this issue. Using this benchmark, models will learn to both admit uncertainty and ground their answers in sources.

Implications for RAG reliability

It is evident that in the evaluated models, poor retrieval does not consistently propagate to the generation stage. For instance, if a retriever pulls obviously unreliable sources (to the human eye), the generator produces high-confidence answers with low hedging. This disconnect makes RAG pipelines seem trustworthy even when a query involves contradictory or sparse evidence.

This has critical implications, especially in domains such as healthcare and finance, where users rely heavily on linguistic cues to make decisions and gauge credibility. We recommend that future RAG systems directly integrate uncertainty-awareness mechanisms in both the retrieval and generation processes. For example, confidence-adjusted retrieval weighting, allowing generators to reflect retrieval source quality more faithfully could be a possible implementation step. Ultimately, a model should develop the ability to communicate its epistemic limits transparently.

Conclusion

We proposed **BLUFF-1000**, a benchmark constructed to evaluate how LLMs manage linguistic confidence under imperfect retrieval conditions. Along with various other metrics for answer correctness, faithfulness, refusals, and numerical overconfidence, we provide a framework for evaluating generation components of RAG systems. We evaluate modern LLMs using gathered sources, varying the source sets provided to models for each question. Through these methods, our most important finding was the consistent patterns of misaligned hedging.

After conducting a controlled experiment on the generation aspect, our findings indicated that future progress in full RAG pipelines as a whole must include developments in uncertainty-aware methods to transparently convey confidence throughout the system. This benchmark provides a foundation for measuring and eventually improving this aspect of model trustworthiness, which is critical for LLMs that serve the user.

References

[1] Ad Fontes Media. 2025. Media Bias Chart® Version 7.0. Accessed October 13, 2025.

[2] Barron, R. C.; Grantcharov, V.; Wanna, S.; Eren, M. E.; Bhattarai, M.; Solovyev, N.; Tompkins, G.; Nicholas, C.; Rasmussen, K. ; Matuszek, C.; and Alexandrov, B. S. 2024. Domain-Specific Retrieval-Augmented Generation Using Vector Stores, Knowledge Graphs, and Tensor Factorization. arXiv:2410.02721.

[3] Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

[4] Cao, T.; Bhandari, N.; Yerukola, A.; Asai, A.; and Sap, M. 2025. Out of Style: RAG's Fragility to Linguistic Variation. arXiv:2504.08231.

[5] Cattan, A.; Jacovi, A.; Ram, O.; Herzig, J.; Aharoni, R.; Goldshtein, S.; Ofek, E.; Szpektor, I.; and Caciularu, A. 2025. DRAGged into Conflicts: Detecting and Addressing Conflicting Sources in Search-Augmented LLMs. arXiv:2506.08500.

[6] Chen, G.; Yao, Y.; Chao, L. S.; Liu, X.; and Wong, D. F. 2025. SGIC: A Self-Guided Iterative Calibration Framework for RAG. arXiv:2506.16172.

[7] DeepSeek-AI; Liu, A.; Feng, B.; Wang, B.; Wang, B.; Liu, B.; Zhao, C.; Dengr, C.; Ruan, C.; Dai, D.; Guo, D.; Yang, D.; Chen, D.; Ji, D.; Li, E.; Lin, F.; Luo, F.; Hao, G.; Chen, G.; Li, G.; Zhang, H.; Xu, H.; Yang, H.; Zhang, H.; Ding, H.; Xin, H.; Gao, H.; Li, H.; Qu, H.; Cai, J. L.; Liang, J.; Guo, J.; Ni, J.; Li, J.; Chen, J.; Yuan, J.; Qiu, J.; Song, J.; Dong, K.; Gao, K.; Guan, K.; Wang, L.; Zhang, L.; Xu, L.; Xia, L.; Zhao, L.; Zhang, L.; Li, M.; Wang, M.; Zhang, M.; Zhang, M.; Tang, M.; Li, M.; Tian, N.; Huang, P.; Wang, P.; Zhang, P.; Zhu, Q.; Chen, Q.; Du, Q.; Chen, R. J.; Jin, R. L.; Ge, R.; Pan, R.; Xu, R.; Chen, R.; Li, S. S.; Lu, S.; Zhou, S.; Chen, S.; Wu, S.; Ye, S.; Ma, S.; Wang, S.; Zhou, S.; Yu, S.; Zhou, S.; Zheng, S.; Wang, T.; Pei, T.; Yuan, T.; Sun, T.; Xiao, W. L.; Zeng, W.; An, W.; Liu, W.; Liang, W.; Gao, W.; Zhang, W.; Li, X. Q.; Jin, X.; Wang, X.; Bi, X.; Liu, X.; Wang, X.; Shen, X.; Chen, X.; Chen, X.; Nie, X.; Sun, X.; Wang, X.; Liu, X.; Xie, X.; Yu, X.; Song, X.; Zhou, X.; Yang, X.; Lu, X.; Su, X.; Wu, Y.; Li, Y. K.; Wei, Y. X.; Zhu, Y. X.; Xu, Y.; Huang, Y.; Li, Y.; Zhao, Y.; Sun, Y.; Li, Y.; Wang, Y.; Zheng, Y.; Zhang, Y.; Xiong, Y.; Zhao, Y.; He, Y.; Tang, Y.; Piao, Y.; Dong, Y.; Tan, Y.; Liu, Y.; Wang, Y.; Guo, Y.; Zhu, Y.; Wang, Y.; Zou, Y.; Zha, Y.; Ma, Y.; Yan, Y.; You, Y.; Liu, Y.; Ren, Z. Z.; Ren, Z.; Sha, Z.; Fu, Z.; Huang, Z.; Zhang, Z.; Xie, Z.; Hao, Z.; Shao, Z.; Wen, Z.; Xu, Z.; Zhang, Z.; Li, Z.; Wang, Z.; Gu, Z.; Li, Z.; and Xie, Z. 2024. DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model. arXiv:2405.04434.

[8] Es, S.; James, J.; Espinosa-Anke, L.; and Schockaert, S. 2025. Ragas: Automated Evaluation of Retrieval Augmented Generation. arXiv:2309.15217.

[9] Golding, B. 2020. Iffy Index of Unreliable Sources.

[10] Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; Yang, A.; Fan, A.; Goyal, A.; Hartshorn, A.; Yang, A.; Mitra, A.; Sravankumar, A.; Korenev, A.; Hinsvark, A.; Rao, A.; Zhang, A.; Rodriguez, A.; Gregerson, A.; Spataru, A.; Roziere, B.; Biron, B.; Tang, B.; Chern, B.; Caucheteux, C.; Nayak, C.; Bi, C.; Marra, C.; McConnell, C.; Keller, C.; Touret, C.; Wu, C.; Wong, C.; Ferrer, C. C.; Nikolaidis, C.; Allonsius, D.; Song, D.; Pintz, D.; Livshits, D.; Wyatt, D.; Esiobu, D.; Choudhary, D.; Mahajan, D.; Garcia-Olano, D.; Perino, D.; Hupkes, D.; Lakomkin, E.; AlBadawy, E.; Lobanova, E.; Dinan, E.; Smith, E. M.; Radenovic, F.; Guzmán, F.; Zhang, F.; Synnaeve, G.; Lee, G.; Anderson, G. L.; Thattai, G.; Nail, G.; Mialon, G.; Pang, G.; Cucurell, G.; Nguyen, H.; Korevaar, H.; Xu, H.; Touvron, H.; Zarov, I.; Ibarra, I. A.; Kloumann, I.; Misra, I.; Evtimov, I.; Zhang, J.; Copet, J.; Lee, J.; Geffert, J.; Vranes, J.; Park, J.; Mahadeokar, J.; Shah, J.; van der Linde, J.; Billock, J.; Hong, J.; Lee, J.; Fu, J.; Chi, J.; Huang, J.; Liu, J.; Wang, J.; Yu, J.; Bitton, J.; Spisak, J.; Park, J.; Rocca, J.; Johnstun, J.; Saxe, J.; Jia, J.; Alwala, K. V.; Prasad, K.; Upasani, K.; Plawiak, K.; Li, K.; Heafield, K.; Stone, K.; El-Arini, K.; Iyer, K.; Malik, K.; Chiu, K.; Bhalla, K.; Lakhotia, K.; Rantala-Yeary, L.; van der Maaten, L.; Chen, L.; Tan, L.; Jenkins, L.; Martin, L.; Madaan, L.; Malo, L.; Blecher, L.; Landzaat, L.; de Oliveira, L.; Muzzi, M.; Pasupuleti, M.; Singh, M.; Paluri, M.; Kardas, M.; Tsimpoukelli, M.; Oldham, M.; Rita, M.; Pavlova, M.; Kambadur, M.; Lewis, M.; Si, M.; Singh, M. K.; Hassan, M.; Goyal, N.; Torabi, N.;

Bashlykov, N.; Bogoychev, N.; Chatterji, N.; Zhang, N.; Duchenne, O.; Çelebi, O.; Alrassy, P.; Zhang, P.; Li, P.; Vasic, P.; Weng, P.; Bhargava, P.; Dubal, P.; Krishnan, P.; Koura, P. S.; Xu, P.; He, Q.; Dong, Q.; Srinivasan, R.; Ganapathy, R.; Calderer, R.; Cabral, R. S.; Stojnic, R.; Raileanu, R.; Maheswari, R.; Girdhar, R.; Patel, R.; Sauvestre, R.; Polidoro, R.; Sumbaly, R.; Taylor, R.; Silva, R.; Hou, R.; Wang, R.; Hosseini, S.; Chennabasappa, S.; Singh, S.; Bell, S.; Kim, S. S.; Edunov, S.; Nie, S.; Narang, S.; Raparthy, S.; Shen, S.; Wan, S.; Bhosale, S.; Zhang, S.; Vandenhende, S.; Batra, S.; Whitman, S.; Sootla, S.; Collot, S.; Gururangan, S.; Borodinsky, S.; Herman, T.; Fowler, T.; Sheasha, T.; Georgiou, T.; Scialom, T.; Speckbacher, T.; Mihaylov, T.; Xiao, T.; Karn, U.; Goswami, V.; Gupta, V.; Ramanathan, V.; Kerkez, V.; Gonguet, V.; Do, V.; Vogeti, V.; Albiero, V.; Petrovic, V.; Chu, W.; Xiong, W.; Fu, W.; Meers, W.; Martinet, X.; Wang, X.; Wang, X.; Tan, X. E.; Xia, X.; Xie, X.; Jia, X.; Wang, X.; Goldschlag, Y.; Gaur, Y.; Babaei, Y.; Wen, Y.; Song, Y.; Zhang, Y.; Li, Y.; Mao, Y.; Coudert, Z. D.; Yan, Z.; Chen, Z.; Papakipos, Z.; Singh, A.; Srivastava, A.; Jain, A.; Kelsey, A.; Shajnfeld, A.; Gangidi, A.; Victoria, A.; Goldstand, A.; Menon, A.; Sharma, A.; Boesenberg, A.; Baevski, A.; Feinstein, A.; Kallet, A.; Sangani, A.; Teo, A.; Yunus, A.; Lupu, A.; Alvarado, A.; Caples, A.; Gu, A.; Ho, A.; Poulton, A.; Ryan, A.; Ramchandani, A.; Dong, A.; Franco, A.; Goyal, A.; Saraf, A.; Chowdhury, A.; Gabriel, A.; Bharambe, A.; Eisenman, A.; Yazdan, A.; James, B.; Maurer, B.; Leonhardi, B.; Huang, B.; Loyd, B.; Paola, B. D.; Paranjape, B.; Liu, B.; Wu, B.; Ni, B.; Hancock, B.; Wasti, B.; Spence, B.; Stojkovic, B.; Gamido, B.; Montalvo, B.; Parker, C.; Burton, C.; Mejia, C.; Liu, C.; Wang, C.; Kim, C.; Zhou, C.; Hu, C.; Chu, C.-H.; Cai, C.; Tindal, C.; Feichtenhofer, C.; Gao, C.; Civin, D.; Beaty, D.; Kreymer, D.; Li, D.; Adkins, D.; Xu, D.; Testuggine, D.; David, D.; Parikh, D.; Liskovich, D.; Foss, D.; Wang, D.; Le, D.; Holland, D.; Dowling, E.; Jamil, E.; Montgomery, E.; Presani, E.; Hahn, E.; Wood, E.; Le, E.-T.; Brinkman, E.; Arcaute, E.; Dunbar, E.; Smothers, E.; Sun, F.; Kreuk, F.; Tian, F.; Kokkinos, F.; Ozgenel, F.; Caggioni, F.; Kanayet, F.; Seide, F.; Florez, G. M.; Schwarz, G.; Badeer, G.; Swee, G.; Halpern, G.; Herman, G.; Sizov, G.; Guangyi; Zhang; Lakshminarayanan, G.; Inan, H.; Shojanazeri, H.; Zou, H.; Wang, H.; Zha, H.; Habeeb, H.; Rudolph, H.; Suk, H.; Aspegren, H.; Goldman, H.; Zhan, H.; Damlaj, I.; Molybog, I.; Tufanov, I.; Leontiadis, I.; Veliche, I.-E.; Gat, I.; Weissman, J.; Geboski, J.; Kohli, J.; Lam, J.; Asher, J.; Gaya, J.-B.; Marcus, J.; Tang, J.; Chan, J.; Zhen, J.; Reizenstein, J.; Teboul, J.; Zhong, J.; Jin, J.; Yang, J.; Cummings, J.; Carvill, J.; Shepard, J.; McPhie, J.; Torres, J.; Ginsburg, J.; Wang, J.; Wu, K.; U, K. H.; Saxena, K.; Khandelwal, K.; Zand, K.; Matosich, K.; Veeraraghavan, K.; Michelena, K.; Li, K.; Jagadeesh, K.; Huang, K.; Chawla, K.; Huang, K.; Chen, L.; Garg, L.; A, L.; Silva, L.; Bell, L.; Zhang, L.; Guo, L.; Yu, L.; Moshkovich, L.; Wehrstedt, L.; Khabsa, M.; Avalani, M.; Bhatt, M.; Mankus, M.; Hasson, M.; Lennie, M.; Reso, M.; Groshev, M.; Naumov, M.; Lathi, M.; Keneally, M.; Liu, M.; Seltzer, M. L.; Valko, M.; Restrepo, M.; Patel, M.; Vyatskov, M.; Samvelyan, M.; Clark, M.; Macey, M.; Wang, M.; Hermoso, M. J.; Metanat, M.; Rastegari, M.; Bansal, M.; Santhanam, N.; Parks, N.; White, N.; Bawa, N.; Singhal, N.; Egebo, N.; Usunier, N.; Mehta, N.; Laptev, N. P.; Dong, N.; Cheng, N.; Chernoguz, O.; Hart, O.; Salpekar, O.; Kalinli, O.; Kent, P.; Parekh, P.; Saab, P.; Balaji, P.; Rittner, P.; Bontrager, P.; Roux, P.; Dollar, P.; Zvyagina, P.; Ratanchandani, P.; Yuvraj, P.; Liang, Q.; Alao, R.; Rodriguez, R.; Ayub, R.; Murthy, R.; Nayani, R.; Mitra, R.; Parthasarathy, R.; Li, R.; Hogan, R.; Battey, R.; Wang, R.; Howes, R.; Rinott, R.; Mehta, S.; Siby, S.; Bondu, S. J.; Datta, S.; Chugh, S.; Hunt, S.; Dhillon, S.; Sidorov, S.; Pan, S.; Mahajan, S.; Verma, S.; Yamamoto, S.; Ramaswamy, S.; Lindsay, S.; Lindsay, S.; Feng, S.; Lin, S.; Zha, S. C.; Patil, S.; Shankar, S.; Zhang, S.; Zhang, S.; Wang, S.; Agarwal, S.; Sajuyigbe, S.; Chintala, S.; Max, S.; Chen, S.; Kehoe, S.; Satterfield, S.; Govindaprasad, S.; Gupta, S.; Deng, S.; Cho, S.; Virk, S.; Subramanian, S.; Choudhury, S.; Goldman, S.; Remez, T.; Glaser, T.; Best, T.; Koehler, T.; Robinson, T.; Li, T.; Zhang, T.; Matthews, T.; Chou, T.; Shaked, T.; Vontimitta, V.; Ajayi, V.; Montanez, V.; Mohan, V.; Kumar, V. S.; Mangla, V.; Ionescu, V.; Poenaru, V.; Mihailescu, V. T.; Ivanov, V.; Li, W.; Wang, W.; Jiang, W.; Bouaziz, W.; Constable, W.; Tang, X.; Wu, X.; Wang, X.; Wu, X.; Gao, X.; Kleinman, Y.; Chen, Y.; Hu, Y.; Jia, Y.; Qi, Y.; Li, Y.; Zhang, Y.; Zhang, Y.; Adi, Y.; Nam, Y.; Yu; Wang; Zhao, Y.; Hao, Y.; Qian, Y.; Li, Y.; He, Y.; Rait, Z.; DeVito, Z.; Rosnbrick, Z.; Wen, Z.; Yang, Z.; Zhao, Z.; and Ma, Z. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

[11] Groot, T.; and Valdenegro-Toro, M. 2024. Overconfidence is Key: Verbalized Uncertainty Evaluation in Large Language and Vision-Language Models. arXiv:2405.02917.

[12] Hwang, J.; Park, J.; Park, H.; Kim, D.; Park, S.; and Ok, J. 2025. Retrieval-Augmented Generation with Estimation of Source Reliability. arXiv:2410.22954.

[13] Jang, C.; Cho, D.; Lee, S.; Lee, H.; and Lee, J. 2025. Reliable Decision Making via Calibration Oriented Retrieval Augmented Generation. arXiv:2411.08891.

[14] Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. arXiv:2310.06825.

[15] Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Hatfield-Dodds, Z.; DasSarma, N.; Tran-Johnson, E.; Johnston, S.; El-Showk, S.; Jones, A.; Elhage, N.; Hume, T.; Chen, A.; Bai, Y.; Bowman, S.; Fort, S.; Ganguli, D.; Hernandez, D.; Jacobson, J.; Kernion, J.; Kravec, S.; Lovitt,

L.; Ndousse, K.; Olsson, C.; Ringer, S.; Amodei, D.; Brown, T.; Clark, J.; Joseph, N.; Mann, B.; McCandlish, S.; Olah, C.; and Kaplan, J. 2022. Language Models (Mostly) Know What They Know. arXiv:2207.05221.

[16] Kalai, A. T.; Nachum, O.; Vempala, S. S.; and Zhang, E. 2025. Why Language Models Hallucinate. arXiv:2509.04664.

[17] Krishna, S.; Krishna, K.; Mohananey, A.; Schwarcz, S.; Stambler, A.; Upadhyay, S.; and Faruqui, M. 2025. Fact, Fetch, and Reason: A Unified Evaluation of Retrieval-Augmented Generation. arXiv:2409.12941.

[18] Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; tau Yih, W.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401.

[19] Li, Z.; Xiong, J.; Ye, F.; Zheng, C.; Wu, X.; Lu, J.; Wan, Z.; Liang, X.; Li, C.; Sun, Z.; Kong, L.; and Wong, N. 2024. UncertaintyRAG: Span-Level Uncertainty Enhanced Long-Context Modeling for Retrieval-Augmented Generation. arXiv:2410.02719.

[20] OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; Avila, R.; Babuschkin, I.; Balaji, S.; Balcom, V.; Baltescu, P.; Bao, H.; Bavarian, M.; Belgum, J.; Bello, I.; Berdine, J.; Bernadett-Shapiro, G.; Berner, C.; Bogdonoff, L.; Boiko, O.; Boyd, M.; Brakman, A.-L.; Brockman, G.; Brooks, T.; Brundage, M.; Button, K.; Cai, T.; Campbell, R.; Cann, A.; Carey, B.; Carlson, C.; Carmichael, R.; Chan, B.; Chang, C.; Chantzis, F.; Chen, D.; Chen, S.; Chen, R.; Chen, J.; Chen, M.; Chess, B.; Cho, C.; Chu, C.; Chung, H. W.; Cummings, D.; Currier, J.; Dai, Y.; Decareaux, C.; Degry, T.; Deutsch, N.; Deville, D.; Dhar, A.; Dohan, D.; Dowling, S.; Dunning, S.; Ecoffet, A.; Eleti, A.; Eloundou, T.; Farhi, D.; Fedus, L.; Felix, N.; Fishman, S. P.; Forte, J.; Fulford, I.; Gao, L.; Georges, E.; Gibson, C.; Goel, V.; Gogineni, T.; Goh, G.; Gontijo-Lopes, R.; Gordon, J.; Grafstein, M.; Gray, S.; Greene, R.; Gross, J.; Gu, S. S.; Guo, Y.; Hallacy, C.; Han, J.; Harris, J.; He, Y.; Heaton, M.; Heidecke, J.; Hesse, C.; Hickey, A.; Hickey, W.; Hoeschele, P.; Houghton, B.; Hsu, K.; Hu, S.; Hu, X.; Huizinga, J.; Jain, S.; Jain, S.; Jang, J.; Jiang, A.; Jiang, R.; Jin, H.; Jin, D.; Jomoto, S.; Jonn, B.; Jun, H.; Kaftan, T.; Łukasz Kaiser; Kamali, A.; Kanitscheider, I.; Keskar, N. S.; Khan, T.; Kilpatrick, L.; Kim, J. W.; Kim, C.; Kim, Y.; Kirchner, J. H.; Kiros, J.; Knight, M.; Kokotajlo, D.; Łukasz Kondraciuk; Kondrich, A.; Konstantinidis, A.; Kosic, K.; Krueger, G.; Kuo, V.; Lampe, M.; Lan, I.; Lee, T.; Leike, J.; Leung, J.; Levy, D.; Li, C. M.; Lim, R.; Lin, M.; Lin, S.; Litwin, M.; Lopez, T.; Lowe, R.; Lue, P.; Makanju, A.; Malfacini, K.; Manning, S.; Markov, T.; Markovski, Y.; Martin, B.; Mayer, K.; Mayne, A.; McGrew, B.; McKinney, S. M.; McLeavey, C.; McMillan, P.; McNeil, J.; Medina, D.; Mehta, A.; Menick, J.; Metz, L.; Mishchenko, A.; Mishkin, P.; Monaco, V.; Morikawa, E.; Mossing, D.; Mu, T.; Murati, M.; Murk, O.; Mély, D.; Nair, A.; Nakano, R.; Nayak, R.; Neelakantan, A.; Ngo, R.; Noh, H.; Ouyang, L.; O'Keefe, C.; Pachocki, J.; Paino, A.; Palermo, J.; Pantuliano, A.; Parascandolo, G.; Parish, J.; Parparita, E.; Passos, A.; Pavlov, M.; Peng, A.; Perelman, A.; de Avila Belbute Peres, F.; Petrov, M.; de Oliveira Pinto, H. P.; Michael; Pokorny; Pokrass, M.; Pong, V. H.; Powell, T.; Power, A.; Power, B.; Proehl, E.; Puri, R.; Radford, A.; Rae, J.; Ramesh, A.; Raymond, C.; Real, F.; Rimbach, K.; Ross, C.; Rotsted, B.; Roussez, H.; Ryder, N.; Saltarelli, M.; Sanders, T.; Santurkar, S.; Sastry, G.; Schmidt, H.; Schnurr, D.; Schulman, J.; Selsam, D.; Sheppard, K.; Sherbakov, T.; Shieh, J.; Shoker, S.; Shyam, P.; Sidor, S.; Sigler, E.; Simens, M.; Sitkin, J.; Slama, K.; Sohl, I.; Sokolowsky, B.; Song, Y.; Staudacher, N.; Such, F. P.; Summers, N.; Sutskever, I.; Tang, J.; Tezak, N.; Thompson, M. B.; Tillet, P.; Tootoonchian, A.; Tseng, E.; Tuggle, P.; Turley, N.; Tworek, J.; Uribe, J. F. C.; Vallone, A.; Vijayvergiya, A.; Voss, C.; Wainwright, C.; Wang, J. J.; Wang, A.; Wang, B.; Ward, J.; Wei, J.; Weinmann, C.; Welihinda, A.; Welinder, P.; Weng, J.; Weng, L.; Wiethoff, M.; Willner, D.; Winter, C.; Wolrich, S.; Wong, H.; Workman, L.; Wu, S.; Wu, J.; Wu, M.; Xiao, K.; Xu, T.; Yoo, S.; Yu, K.; Yuan, Q.; Zaremba, W.; Zellers, R.; Zhang, C.; Zhang, M.; Zhao, S.; Zheng, T.; Zhuang, J.; Zhuk, W.; and Zoph, B. 2024. GPT-4 Technical Report. arXiv:2303.08774.

[21] Ouyang, J.; Pan, T.; Cheng, M.; Yan, R.; Luo, Y.; Lin, J.; and Liu, Q. 2025. HoH: A Dynamic Benchmark for Evaluating the Impact of Outdated Information on Retrieval-Augmented Generation. arXiv:2503.04800.

[22] Ozaki, S.; Kato, Y.; Feng, S.; Tomita, M.; Hayashi, K.; Hashimoto, W.; Obara, R.; Oyamada, M.; Hayashi, K.; Kamigaito, H.; and Watanabe, T. 2025. Understanding the Impact of Confidence in Retrieval Augmented Generation: A Case Study in the Medical Domain. arXiv:2412.20309.

[23] Pipitone, N.; and Alami, G. H. 2024. LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain. arXiv:2408.10343.

[24] Ru, D.; Qiu, L.; Hu, X.; Zhang, T.; Shi, P.; Chang, S.; Jiayang, C.; Wang, C.; Sun, S.; Li, H.; Zhang, Z.; Wang, B.; Jiang, J.; He, T.; Wang, Z.; Liu, P.; Zhang, Y.; and Zhang, Z. 2024. RAGChecker: A Fine-grained Framework for Diagnosing Retrieval-Augmented Generation. arXiv:2408.08067.

[25] Semnani, S. J.; Burapacheep, J.; Khatua, A.; Atchariyachanvanit, T.; Wang, Z.; and Lam, M. S. 2025. Detecting Corpus-Level Knowledge Inconsistencies in Wikipedia with Large Language Models. arXiv:2509.23233.

[26] Siriwardhana, S.; Weerasekera, R.; Wen, E.; Kaluarachchi, T.; Rana, R.; and Nanayakkara, S. 2022. Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering. arXiv:2210.02627.

[27] Sorodoc, I.-T.; Ribeiro, L. F. R.; Blloshmi, R.; Davis, C.; and de Gispert, A. 2025. GaRAGe: A Benchmark with Grounding Annotations for RAG Evaluation. arXiv:2506.07671.

[28] Sun, H.; Wang, Y.; and Zhang, S. 2024. Retrieval-Augmented Generation for Domain-Specific Question Answering: A Case Study on Pittsburgh and CMU. arXiv:2411.13691.

[29] Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *Science*, 359(6380): 1146–1151.

[30] Wang, F.; Wan, X.; Sun, R.; Chen, J.; and Arik, S. O. 2025. Astute RAG: Overcoming Imperfect Retrieval Augmentation and Knowledge Conflicts for Large Language Models. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 30553–30571. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

[31] Yang, A.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Huang, H.; Jiang, J.; Tu, J.; Zhang, J.; Zhou, J.; Lin, J.; Dang, K.; Yang, K.; Yu, L.; Li, M.; Sun, M.; Zhu, Q.; Men, R.; He, T.; Xu, W.; Yin, W.; Yu, W.; Qiu, X.; Ren, X.; Yang, X.; Li, Y.; Xu, Z.; and Zhang, Z. 2025. Qwen2.5-1M Technical Report. arXiv:2501.15383.

[32] Yang, X.; Sun, K.; Xin, H.; Sun, Y.; Bhalla, N.; Chen, X.; Choudhary, S.; Gui, R. D.; Jiang, Z. W.; Jiang, Z.; Kong, L.; Moran, B.; Wang, J.; Xu, Y. E.; Yan, A.; Yang, C.; Yuan, E.; Zha, H.; Tang, N.; Chen, L.; Scheffer, N.; Liu, Y.; Shah, N.; Wanga, R.; Kumar, A.; tau Yih, W.; and Dong, X. L. 2024. CRAG – Comprehensive RAG Benchmark. arXiv:2406.04744.

Appendix

Code and data for BLUFF-1000 are open source and will be available after reviewing process.

Handling Noisy Environments

RA-RAG [12] and CLAIRE [25] were developed to include the evaluation of retrieved sources to ensure robust and factual responses. Retrieved source evaluation included source reliability estimation and knowledge inconsistency detection. Furthermore, to combat the conflicts that may arise from combining internal and external information, Astute RAG was proposed as a novel RAG approach [30]. While accounting for source reliability, internal and external varieties of information are extracted and combined. Subsequently, a final response is produced based on their assessed trustworthiness.

Question-Answer Pair Generation Outline

First, a set of possible subdomains for each domain was formulated. (See Table 2 in Appendix) After, the 4o model was prompted to create a unique query under a given subdomain, which was then sent to the Duck Duck Go (DDG) API, where only sources established as reliable were selected as gold sources. (See "Source Filtering" for this process) Then, based on the gold source content, the model was prompted to create question and answer. Source metadata was also stored. This question generation process was repeated 500 times.

Source Filtering

To select sources whose reliability had already been established, a comprehensive catalog of sources and their classifications as reliable or unreliable was assembled. When compiling the source list, named the Domain Trust Configuration, we referenced pre-researched records of source reliability. The Ad Fontes Media Bias List gathers political analysts to quantify source factuality and bias [1], and the Iffy Index of Reliable Sources [9], developed by The University of Michigan Center for Social Media Responsibility, provides a comprehensive collection of websites that regularly publish untrue information. We combined all sources and classifications from both sets to construct the Domain Trust Configuration. Additionally, all top-level domains ".gov" or ".edu" were classified as reliable.

The second round of filtering ensured that the dataset consisted only of sources relevant to the gold question. Relevancy scores were calculated by measuring the frequency of common keywords between the gold questions and source texts. Only sources with relevancy scores above a 0.3 threshold were added to our final dataset. This process minimized the number of unrelated sources that may have emerged during the source gathering process.

Obtaining Sources Outline

After question generation, sources were appended to each question to simulate retrieval. For each dataset entry, search queries were created by appending reliable keywords (e.g "government", "research", etc.) to the end of each gold question. Afterward, this process was repeated with unreliable keywords (e.g. "conspiracy", "hoax", etc.). The top search results were gathered, and the results were refined to include only relevant sources with predefined reliability classifications (See "Source Filtering" in Appendix).

For all questions, clear source sets were primarily composed of reliable sources, while ambiguous sets contained primarily unreliable sources. (See Figures 6 % 7 for the format of source sets)

If the initial search did not yield enough unreliable sources, the dataset generation script turned to public forums, specifically Reddit, due to the inherently unreliable nature of these posts. Unreliable sources were chosen from the most popular comments in related Subreddits.

2 distraction sources unrelated to the questions were also included ambiguous source set of each question to evaluate the ability of models to reject irrelevant information. (Note: distraction sources do not have a relevancy score; See Figure 10 in the Appendix for a visualization of the source gathering process).

Metric Definitions

Lexical Overconfidence Index measures the overconfident language in incorrect answers.

Refusal Rate measures the proportion of questions where the confidence of the model drops below the 20th percentile threshold, causing it to refuse to answer.

Refusal Sensitivity measures the difference in refusal rate of the model between ambiguous and clear information sets.

Source Set On Hedging measures the change in hedging rate between clear and ambiguous questions throughout the dataset.

Source Awareness Score measures the correlation between model confidence and source quality by comparing model confidence under clear and ambiguous sets of evidence as well as analyzing diversity of sources to determine the source quality.

Hedge Precision measures the proportion of hedges used appropriately when the model is incorrect.

Hedge Recall measures the proportion of incorrect answers in which hedging was used.

Answer Correctness measures the average correctness across all answered questions. Done with GPT-4o grader.

Faithfulness evaluates an LLM's ability to stay faithful to information from the sources it retrieves with RAG while avoiding the hallucination of information. Calculated using *Answer-Source Overlap*, *Hallucination Rate (HR)*, *Attribution Coverage (AC)*, and *Grounding Score (GS)*.

Metric Formulas

$$\text{Lexical Overconfidence} = \frac{1}{|\text{Wrong}|} \sum_{i \in \text{Wrong}} \frac{\text{confident\_terms}_i}{\text{word\_count}_i}$$

$$\text{Refusal Rate} = \frac{1}{n} \sum_{i=1}^{n} 1[c_i < c_{\text{threshold}}]$$

$$\text{Refusal Sensitivity} = \text{Ambig. Refusal Rate} - \text{Clear Refusal Rate}$$

$$\text{SourceSetOnHedging} = \frac{1}{|R_a|} \sum_{r \in R_a} h(r) - \frac{1}{|R_c|} \sum_{r \in R_c} h(r)$$

where R_a the set of responses with ambiguous source sets. R_c is the set of responses with clear source sets. h(r) is the hedge count function that counts hedge terms in response.

$$\text{Source Awareness Score} = \text{Corr}(\text{Source Quality}, \text{Confidence})$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Answer Correctness} = \frac{1}{n} \sum_{i=1}^{n} a_i$$

$$\text{Overlap} = \frac{|\text{Tokens}(\text{Answer}) \cap \text{Tokens}(\text{Sources})|}{|\text{Tokens}(\text{Answer}) \cup \text{Tokens}(\text{Sources})|}$$

$$\text{HR} = \frac{\text{Unsupported Claims}}{\text{Total Claims}}$$

$$\text{AC} = \frac{\text{Attributed Claims}}{\text{Total Claims}}$$

$$\textbf{Faithfulness} = 0.3 \times \text{Overlap} + 0.2 \times (1 - \text{HR}) + 0.3 \times \text{AC} + 0.2 \times \text{GS}$$

$$\text{GS} = \frac{1}{|\text{Sources}|} \sum_{s=1}^{|\text{Sources}|} \text{Overlap}(\text{Answer}, \text{Source}_s)$$

$$\text{Corr}(C, A) = \frac{\sum_{i=1}^{n} (c_i - \bar{c})(a_i - \bar{a})}{\sqrt{\sum_{i=1}^{n} (c_i - \bar{c})^2 \sum_{i=1}^{n} (a_i - \bar{a})^2}}$$

```
 {
   "id": "psychology_031",
   "topic": "psychology",
   "question": "How does motivational interviewing differ from traditional directive or
confrontational approaches in addiction therapy?",
   "gold_answer": "Motivational interviewing differs from traditional directive or
confrontational approaches in that…",
   "gold_passage": {
     "title": "www.highfocuscenters.com",
     "url":
"https://www.highfocuscenters.com/education-center/what-is-motivational-interviewing/",
     "date": "2025-08-25",
     "text": "Education Center Motivational Interviewing is a style of counseling…"
   },
```


Figure 5: Sample dataset question (”gold_answer” and ”text” have been shortened for the sake of size)


```
"ambiguous": [
   {
     "url": "https://pmc.ncbi.nlm.nih.gov/articles/PMC6193639/",
     "domain": "pmc.ncbi.nlm.nih.gov",
     "category": "reliable",
     "title": "Effectiveness of Motivational Interviewing on adult behaviour ...",
     "text": "To synthesize the evidence…",
     "timestamp": "2025-08-25T00:47:34.553086",
     "score": 0.39135135135135135
   },
   {
     "url": "https://wellwisp.com/what-is-motivational-interviewing/",
     "domain": "wellwisp.com",
     "category": "unreliable",
     "title": "What Is Motivational Interviewing ? | Transformative Conversations",
     "text": "For instance: - Client: \"I know I should quit smoking, but it's so hard…",
     "timestamp": "2025-08-25T00:47:34.553087",
     "score": 0.4437267080745342
   },
   {
     "url": "https://reddit.com/r/mentalhealth/comments/nwmpz1/im_sick_of_therapists_using_cbt_where_it_isnt/",
     "domain": "reddit.com",
     "category": "unreliable",
     "title": "I'm sick of therapists using CBT where it isn't helpful",
     "text": "I'm sick of therapists..",
     "timestamp": "2025-08-25T00:47:34.553088",
     "score": 0.3824657534246575
   },
   {
     "url": "https://en.m.wikipedia.org/wiki/List_of_effects",
     "domain": "en.m.wikipedia.org",
     "category": "distraction",
     "title": "en.m.wikipedia.org",
     "text": "This is a list of names…",
     "timestamp": "2025-08-25T00:47:34.556035",
     "score": "N/A"
   },
   {
     "url": "https://en.wikipedia.org/wiki/Sleep_deprivation",
     "domain": "en.wikipedia.org",
     "category": "distraction",
     "title": "en.wikipedia.org",
     "text": "Sleep deprivation…",
     "timestamp": "2025-08-25T00:47:34.556056",
     "score": "N/A"
   }
 ]
```


Figure 6: Sample ambiguous source set for a dataset entry (”text” has been shortened for the sake of size)

```
"clear": [
  {
    "url": "https://pmc.ncbi.nlm.nih.gov/articles/PMC3330017/",
    "domain": "pmc.ncbi.nlm.nih.gov",
    "category": "reliable",
    "title": "Motivational Interviewing: moving from why to how with ...",
    "text": "A traditional counseling…",
    "timestamp": "2025-08-25T00:47:34.553068",
    "score": 0.5754098360655737
  },
  {
    "url": "https://pmc.ncbi.nlm.nih.gov/articles/PMC6927483/",
    "domain": "pmc.ncbi.nlm.nih.gov",
    "category": "reliable",
    "title": "Applying Motivational Interviewing Strategies to Enhance ...",
    "text": "16, 18-19 The spirit of motivational interviewing is a…",
    "timestamp": "2025-08-25T00:47:34.553080",
    "score": 0.45104294478527607
  },
  {
    "url": "https://pmc.ncbi.nlm.nih.gov/articles/PMC1463134/",
    "domain": "pmc.ncbi.nlm.nih.gov",
    "category": "reliable",
    "title": "Motivational interviewing: a systematic review and meta- ...",
    "text": "- \u25b8 The therapeutic relationship is a partnership …",
    "timestamp": "2025-08-25T00:47:34.553082",
    "score": 0.5482758620689655
  },
  {
    "url": "https://pmc.ncbi.nlm.nih.gov/articles/PMC8200683/",
    "domain": "pmc.ncbi.nlm.nih.gov",
    "category": "reliable",
    "title": "Motivational Interviewing: An Evidence-Based Approach ...",
    "text": "The goal is to increase personal motivation…",
    "timestamp": "2025-08-25T00:47:34.553084",
    "score": 0.4846153846153846
  },
  {
    "url": "https://reddit.com/r/Psychiatry/comments/1khab9y/do_you_prescribe_any_medication_when_treating/",
    "domain": "reddit.com",
    "category": "unreliable",
    "title": "Do you prescribe any medication when treating cocaine discontinuation?",
    "text": "Do you prescribe any medication…",
    "timestamp": "2025-08-25T00:47:34.553085",
    "score": 0.4616867469879518
  }
],
```


Figure 7: Sample clear source set for a dataset entry (”text” has been shortened for the sake of size)

```
You are an expert assistant tasked with answering questions based on
provided source materials.

Question: {question}

Source Materials:
{sources_text}

Instructions:
1. Answer the question based ONLY(VERY IMPORTANT) on the information
provided in the sources above : DO NOT USE YOUR OWN KNOWLEDGE BASE OR
SEARCH THE WEB.
2. If the sources contain conflicting information, acknowledge this in
your response
3. If the sources don't contain enough information to answer
confidently, say so
4. Use appropriate hedging language when uncertain (e.g., "likely",
"probably", "possibly", etc.)
5. Do not exceed 75 words in your response.
6. Answer only if you are >{confidence_threshold:.1f} confident, if
not, refuse. Since refusals will be awarded 0.4 points and having less
than 0.2 accuracy will be penalized more than if you had refused using
"I don't know" as your answer.
7. There should never be a case where you answer the question brutally
wrong. If you believe you will answer the question wrong(<= 0.2
confidence), refuse with "I don't know" as your answer.
```

Figure 8: Prompt for evaluations, emphasizing refusals

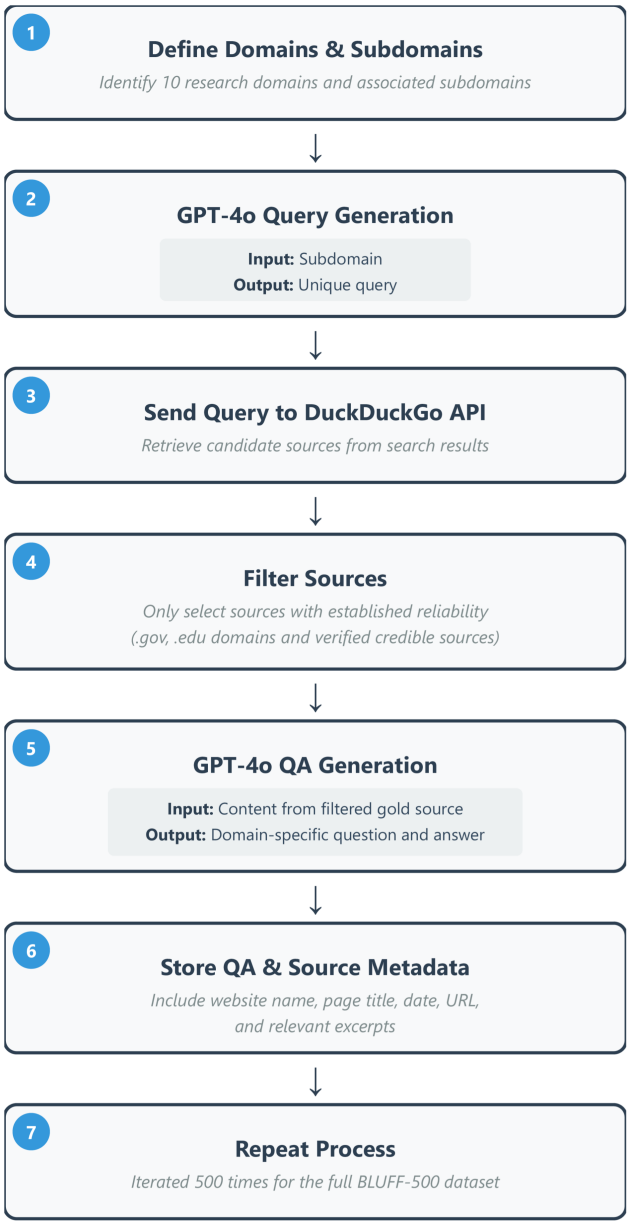

Figure 9: QA Pair Generation Workflow

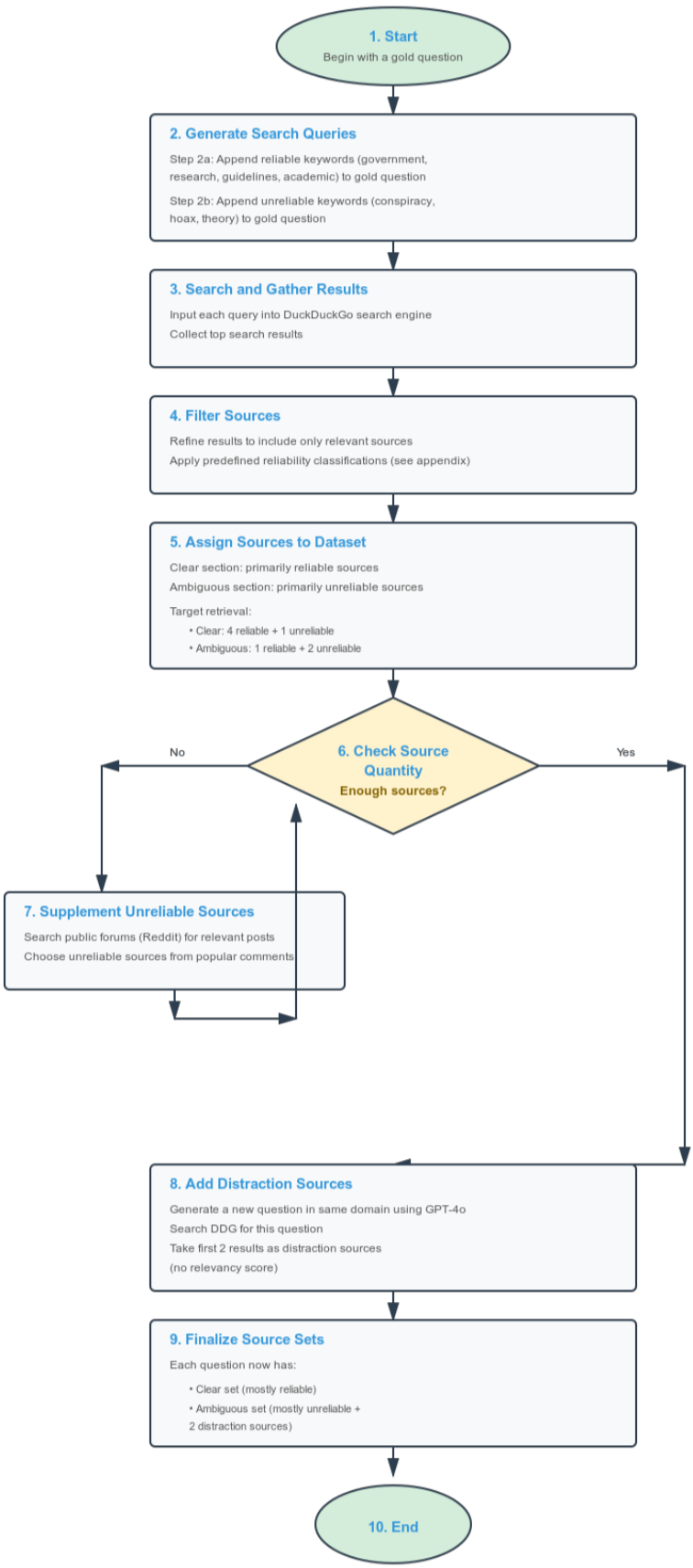

Figure 10: Source Gathering Flowchart

Table 1: Comprehensive Model Performance Comparison

Metric	GPT-4o	GPT-3.5	Llama-4	Llama-3.3	DeepSeek-V3.1	Qwen2.5	Mistral-7B
ASI	-0.024	-0.131	0.067	0.000	-0.017	-0.009	0.018
VUI	0.057	0.054	0.119	0.327	0.120	0.046	0.111
Source Set on Hedging	-0.318	-1.400	0.000	-0.294	0.100	-0.123	-0.050
Lexical Overconfidence	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Hedge Precision	0.029	0.028	0.063	0.198	0.065	0.024	0.059
Hedge Recall	1.000	0.750	0.931	0.941	0.732	0.737	0.903
Refusal Count	380	340	19	152	28	116	0
Refusal Sensitivity	0.206	0.431	-0.038	0.176	0.058	0.202	0.000
Answer Correctness	0.554	0.687	0.683	0.495	0.674	0.785	0.723
Overall Faithfulness	0.647	0.644	0.749	0.722	0.716	0.712	0.717

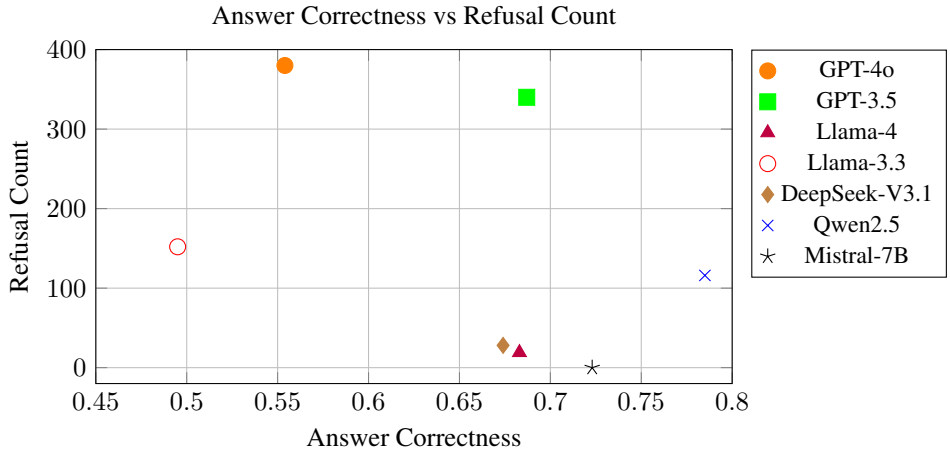

Figure 11: Relationship between answer correctness and refusal count across different models.

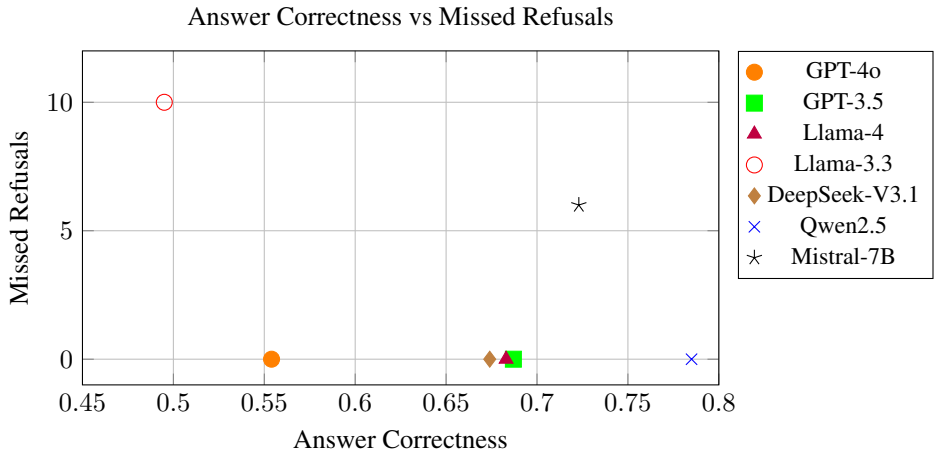

Figure 12: Relationship between answer correctness and missed refusals across different models.

Domain	Subdomain Group 1	Subdomain Group 2	Subdomain Group 3
Astronomy	planets_and_moons	stars_and_galaxies	black_holes
	space_exploration	telescopes	cosmology
	exoplanets	astrobiology	universe_origin
	moon_landing	-	-
Climate	climate_change	carbon_emissions	renewable_energy
	sea_level_rise	deforestation	pollution
	sustainable_development	climate_policy	-
Current Events	international_conflicts	elections	natural_disasters
	pandemics	economic_crises	major_legislation
	scientific_breakthroughs	protests_and_movements	-
Finance	stock_markets	banking	cryptocurrency
	inflation_and_recession	personal_finance	global_trade
	housing_markets	investment_strategies	-
History	ancient_civilizations	world_wars	revolutions
	colonialism	cold_war	civil_rights_movements
	historical_figures	archaeological_discoveries	us-history
Law	constitutional_law	criminal_law	international_law
	civil_rights	supreme_court_cases	intellectual_property
	environmental_law	human_rights_law	immigration_law
Psychology	clinical_psychology	cognitive_psychology	developmental_psychology
	social_psychology	behavioral_psychology	neuropsychology
	forensic_psychology	health_psychology	industrial_psychology
	educational_psychology	abnormal_psychology	personality_psychology
	experimental_psychology	counseling_psychology	sports_psychology
Public Health	vaccines	infectious_diseases	nutrition_guidelines
	mental_health	maternal_and_child_health	chronic_diseases
	occupational_safety	toxicology	emergency_medicine
	preventive_care	healthcare_policy	medical_research
	public_health_education	-	-
Sports	olympics	soccer	basketball
	baseball	tennis	athletics
	sports_medicine	sports_history	doping_scandals
	football	-	-
Technology	artificial_intelligence	social_media	cybersecurity
	5g_networks	biotechnology	quantum_computing
	software_and_internet	consumer_electronics	-

Table 2: Domains and their associated subdomains. Subdomains are grouped across columns to reduce table height and improve readability in a multi-column format

Figure 13: Overconfident Response with Unreliable Sources
Model: Mistral-7B,
Question: How has the methodology for estimating CO2 emissions from the cement industry in China evolved to improve accuracy in recent years?

Model Response: The methodology for estimating CO2 emissions from the cement industry in China has evolved to improve accuracy by incorporating artificial intelligence-based neural networks, specifically Long Short-Term Memory (LSTM) neural networks.

Explanation: This approach is aimed at handling the nonlinear relationships and complex dynamics inherent in emission datasets(Source 1), thereby providing more accurate predictions. In contrast, traditional methods such as statistical analyses and empirical models may not fully capture the long-term dependencies and dynamic changes in CO2 emissions influenced by factors like energy consumption patterns, production processes, and policy interventions.

Gold Answer: The methodology evolved by using provincial clinker production data and provincial clinker emission factors from multiple official statistics sources, following the Intergovernmental Panel on Climate Change (IPCC) methodology, to create a more accurate and high-quality panel dataset for

estimating CO2 emissions.

Source Information:

- **Source 1:** journals.plos.org (reliable) - Estimation of carbon dioxide emissions from the cement industry in Beijing-Tianjin-Hebei using neural networks
- **Source 2:** pubmed.ncbi.nlm.nih.gov (reliable) - Empirical assessing cement CO2 emissions based on China's economic and social development during 2001-2030
- **Source 3:** cdc.gov (distraction) - There are millions of U.S. workers exposed to chemicals in their work each year
- **Source 4:** pmc.ncbi.nlm.nih.gov (distraction) - In many industrial sectors, workers are exposed to manufactured or unintentionally emitted airborne nanoparticles (NPs)
- **Source 5:** reddit.com (unreliable) - Coronavirus is about to trigger a global economic crisis, and the public doesn't care because the media won't publicize it

Performance Metrics:

- **Confidence:** 0.990
- **Accuracy:** 0.400
- **Faithfulness:** 0.855
- **Source Type:** Ambiguous