CITECHECK: Towards Accurate Citation Faithfulness Detection

Anonymous ACL submission

Abstract

Citation faithfulness detection is critical for enhancing retrieval-augmented generation (RAG) systems, yet large-scale Chinese datasets for this task are scarce. Existing methods face prohibitive costs due to the need for manually annotated negative samples. To address this, we introduce the first large-scale Chinese dataset CITECHECK for citation faithfulness detection, constructed via a cost-effective approach using two-stage manual annotation. This method balances positive and negative samples while significantly reducing annotation expenses. CITECHECK comprises training and test splits. Experiments demonstrate that: (1) the test samples are highly challenging, with even state-of-the-art LLMs failing to achieve high accuracy; and (2) training data augmented with LLM-generated negative samples enables smaller models to attain strong performance using parameter-efficient fine-tuning. CITECHECK provides a robust foundation for advancing citation faithfulness detection in Chinese RAG systems.

1 Introduction

Large Language Models (LLMs) are prone to generating factual errors through hallucinations when answering real-world questions. Retrieval-augmented generation (RAG) systems (Lewis et al., 2020; Guu et al., 2020; Borgeaud et al., 2022) address this limitation by leveraging external information retrieval to ground LLM responses in verifiable sources. Recent advancements extend RAG systems to generate text with inline citations (Gao et al., 2023b), enabling users to validate the reliability of generated content by cross-referencing cited documents. However, studies reveal a critical weakness in these systems: citation faithfulness. A substantial portion of generated text may lack proper support from the cited references (Liu et al., 2023; Hu et al., 2024b), undermining the trustworthiness and verification capability of RAG outputs. This challenge necessitates accurate citation faithfulness detection—determining whether cited passages genuinely support their associated claims—as a fundamental requirement for improving RAG reliability.

Developing robust citation faithfulness detection methods requires large-scale, high-quality datasets. While English benchmarks have emerged (Yue et al., 2023), Chinese datasets remain notably absent. Constructing such resources presents unique challenges: realistic negative samples (unsupported citations) from strong RAG systems are usually highly judgmentally difficult and meaningful for studies, yet these systems rarely produce such errors. For instance, a RAG system with a 10% error rate would require annotating approximately 70,000 samples to collect 7,000 negative examples—a prohibitively expensive endeavor. This tension between dataset quality and construction cost demands innovative solutions for efficient data curation without compromising sample integrity.

To bridge this gap, we introduce CITECHECK, the first large-scale Chinese dataset for citation faithfulness detection. Our approach combines 11,307 knowledge-intensive questions with a novel two-stage annotation framework that reduces labeling costs while preserving data quality. CITECHECK comprises two distinct components designed to address both detection difficulty and training efficacy.

The development and test sets each contain 500 positive (supported) and 500 negative (unsupported) samples totaling 2,000 unmodified RAG outputs. Experimental analysis demonstrates these original samples pose significant challenges, with state-of-the-art LLMs achieving limited detection accuracy. The training set includes 9,796 samples (4,898 positive/negative pairs) where negative instances are generated through LLM-based document modification rather than relying solely on rare RAG errors. Despite this augmentation, parameter-

efficient fine-tuning on 7B-8B parameter models yields strong detection performance, confirming the preserved quality of modified negative samples.

Our contributions are threefold: CITECHECK establishes the first comprehensive benchmark for Chinese citation faithfulness detection; (2) We propose an efficient data augmentation strategy that reduces annotation costs by 86% compared to conventional approaches; (3) Extensive experiments validate the dataset's quality and utility, showing that models trained on our augmented data effectively generalize to challenging real-world samples. This work advances reliable RAG development by providing essential resources and methodologies for building verifiable, citation-grounded LLM applications in Chinese.

2 Dataset Construction

2.1 Question Collection

We collect Chinese questions from the sources:
WebText (Xu, 2019): A large-scale Chinese community question-answering dataset spanning diverse topics.

WebCPM (Qin et al., 2023): A Chinese long-form question-answering dataset focused on interactive web search contexts.

Zhihu-KOL (Wang, 2023): A high-quality question-answering dataset derived from Zhihu, a prominent Chinese QA platform.

RGB (Chen et al., 2024): A bilingual question-answering dataset based on news reports.

TrickQA: Questions with ambiguous, incorrect, or unverifiable premises (see Appendix A for details).

After collecting these questions, we input them into an open-sourced RAG system to simulate real-world question-answering scenarios and analyze how the system processes and responds to these diverse inputs. The RAG system retrieves five external documents and generates responses. Statements in the answers are annotated with citation marks (1–5), indicating alignment with information from the corresponding documents. On average, each statement spans 33.4 tokens, while each document averages 177.3 tokens. An original sample is formed by pairing a labeled statement with its cited documents, represented as a tuple (question, answer, statement, cited documents). See Appendix B for more statistics.

2.2 Data Augmentation

The goal of data augmentation is to create negative samples of high quality by making minor modifications to the cited documents in the original samples. Given the use of an industrial RAG system, the number of negative samples in the original samples is estimated to be insufficient. To construct a balanced training set, as well as a label-balanced dev set and test set for evaluation, successfully augmented negative samples can be used. The modified documents should not be inconsistent or incoherent, so as not to provide the trained model with a false basis for judging the negative samples.

We use GPT-4o (OpenAI et al., 2024) for data augmentation. After providing the original sample to the LLM, it is asked to perform the following steps in sequence:

Segments Identification: Find all key segments in the cited document that directly support the information in the statement.

Segments Grouping: Group the key segments by the information they support, with each group containing key segments that support the same or related information in the statement.

Segments Modification: Select a group of key segments and modify them so that they do not support the corresponding information in the statement.

The modification changes only the portion that relates to the supported information in the statement. This maintains logical flow and non-contradictory information within the key segments, and keep the key segments logical in the context of the document and non-contradictory to other information in the document. If there is more than one key segment in a group, the information in all of them should be consistent after the modification.

For each sample, the LLM is asked to try two methods of modification:

Content Revision: Alter specific details within a key segment without introducing direct contradictions to the original information.

Structure Preservation: Remove information from a key segment while ensuring the overall coherence and integrity of the segment remain intact.

After completing the LLM augmentation, each original sample is accompanied by the LLM-labeled key segment information and corresponds to the two augmented samples generated by the LLM using the two modification methods. The cost is 0.026$ per sample. See Appendix C for more details of the augmentation.

Question	特斯拉在中国的纯电动汽车销量占比是多少? What is Tesla's share of all-electric car sales in China?	乘坐飞机的时候托运一个行李箱，再带一个20寸的箱子，带上飞机的箱子会被称重吗? When I check a suitcase on an airplane and bring a 20" case, will the case be weighed when I bring it on the plane?	为什么当你把一幅画切成几块时，每一块都只有一部分，但是当你把一张全息图切成几块时，每一块都有完整的图像? Why is it that when you cut a painting into pieces, each piece has only a part of it, but when you cut a hologram into pieces, each piece has the full image?	如果闪电劈中混凝土大楼之类的，就算在房间里处于隔缘状态，也会受伤害么? If lightning strikes a concrete building or something like that, can you get hurt even if you're in a room that's insulated?
Answer (Statement)	特斯拉在2023年上半年纯电动汽车市场的占有率为21.7%。 Tesla's share of the all-electric car market in the first half of 2023 is 21.7 percent.	是的，乘坐飞机时...。每件托运行李的重量不能超过50公斤，而每位旅客随身携带物品的重量以5公斤为限。此外，每位旅客...。 The weight of each piece of checked baggage cannot exceed 50 kilograms, while the weight of each passenger's carry-on is limited to 5 kilograms.	这是因为...。全息的原理在于将动量空间通过全息技术记录在实空间里，因此，如果全息图缺失一部分，那部分实空间的信息就丢失了，像就不完整了。即使...。 If a portion of the hologram is missing, that portion of real space is lost and the likeness is incomplete.	闪电击中混凝土大楼... 在澳大利亚新南威尔士州纽卡斯尔的一个案例中，闪电击中了一座大楼，虽然大楼的结构保持良好，但是巨大的爆裂声和震动可能会对内部的人造成伤害。因此，...。 The building is structurally sound.
Cited Documents	[1]【2023上半年全球纯电动汽车销量出炉...】...据该报道，特斯拉在纯电动汽车市场期间占据21.7%的份额。... First half of 2023 / Tesla held a 21.7% share of the all-electric car market during the period.	[1] 办理托运行李对行李物品规定如下：...每件行李物品重量不能超过50公斤。... Check-in baggage / The weight of each baggage item can not exceed 50 kilograms. [2] 随身携带物品的重量，每位旅客以5公斤为限。... The weight of carry-on items is limited to 5 kg per passenger.	[1] ...如果普通照片缺失一部分，那部分实空间的信息就丢失了，像就不完整了。全息照片如果缺失一部分，同样会造成信息的缺失，但是... If a portion of an ordinary photograph is missing, that portion of real space is lost and the likeness is incomplete.	[1] ...澳大利亚新南威尔士州纽卡斯尔...可清楚看到闪电击中大楼的场面，同时可听到巨大的爆裂声。据悉，闪电所击中的大楼为一处健身房。...
Label	positive	positive	negative	negative
Note	supported by a single document	supported by multiple documents	contradictory information	unmentioned information

Table 1: Sample examples of the dataset. For the answer and cited documents we show only part of the content. We underline the selected statement in the answer. We mark in red and blue the key information associated with the label in the statement and the cited documents. We provide English translations of the questions and key information.

2.3 Two-stage Manual Annotation

The original samples need to be manually labeled as positive or negative samples before they can be used to form the dataset (examples are shown in Table 1). In the LLM augmentation phase, although we try to guide the LLM to augment negative samples with qualified quality, the LLM may generate some samples that do not meet the requirements. Therefore, the augmented samples also need to be manually labeled for compliance before they can be used to form the dataset. The goal of the two-stage manual annotation is to complete the manual annotation needed above.

In the first stage, the annotators (from the professional data annotation institution in China) need to label whether the original sample is a positive or negative sample, i.e., to determine whether the sum of the information provided by the cited documents fully supports the statement. In order to reduce the difficulty of labeling, the information of key segments labeled by LLM will be provided to the annotators as a reference. However, since the LLM labeling is not always accurate, if the annotators are unable to make a judgment after reading the key segments, they still need to read other parts of the documents to make a judgment. In this stage, the number of negative samples identified by the annotation is 1,006, with a negative sample rate of about 9%. We randomly selected 2,000 samples (1,000 negative and 1,000 positive) and split them equally to create the development and test sets. The augmented samples corresponding to the positive samples in the remaining original samples will be labeled in the second stage.

In the second stage, the annotators need to determine whether an augmented sample is of acceptable quality and whether it is a negative sample. In order to reduce the difficulty of labeling, we show the annotator a comparison of the documents before and after the modification in the form of modification traces. Among the augmented samples that the annotators determine to be negative samples of acceptable quality, we select 2,449 samples that use the modification methods of changing information and deleting information respectively, totaling 4,898 samples. These augmented negative samples together with the 4,898 positive samples in the original samples identified by the first stage of annotation constitute the training set. The two-stage manual annotation costs 0.5$ per sample. See Appendix D for more details on annotation.

3 Experiments

In our experiments, we evaluate the dataset using two approaches. First, we assess the zero-shot performance of state-of-the-art LLMs on the de-

Dev	Acc	Acc_p	Acc_n
GPT-4o	**83.7**	97.0	70.4
Qwen2.5-Plus	81.6	97.0	66.2
DeepSeek-v3	69.4	99.2	39.6
Llama-3.1-8B	**91.4**	91.6	91.2
Mistral-7B	89.5	91.2	87.8
Qwen2.5-7B	91.2	95.0	87.4
Test	**Acc**	Acc_p	Acc_n
GPT-4o	**83.9**	96.2	71.6
Qwen2.5-Plus	81.2	94.8	67.6
DeepSeek-v3	69.4	99.4	39.4
Llama-3.1-8B	**90.6**	90.4	90.8
Mistral-7B	89.8	92.0	87.6
Qwen2.5-7B	88.5	90.4	86.6

Table 2: Results of experiments on the dev set and the test set. We report overall accuracy (**Acc**), accuracy on positive samples ($\textbf{Acc}_\text{p}$), and accuracy on negative samples ($\textbf{Acc}_\text{n}$) in percentage form.

velopment and test sets. This aims to highlight the challenge posed by the test samples. Second, due to resource constraints, we conduct parameter-efficient fine-tuning on smaller models using the training data. This focuses on demonstrating the effectiveness of the training samples. See Appendix F for more experiments for quality validation.

3.1 Settings

State-of-the-art LLMs that we use for zero-shot performance tests include GPT-4o (OpenAI et al., 2024), Qwen2.5-Plus (Qwen et al., 2024), and DeepSeek-v3 (DeepSeek-AI et al., 2024). We provide the sample to the LLMs and ask for their judgment. The relatively small language models we use for training include Llama-3.1-8B (Grattafiori et al., 2024), Mistral-7B (Jiang et al., 2023), and Qwen2.5-7B (Qwen et al., 2024). The parameter-efficient fine-tuning method we use is LoRA (Hu et al., 2022). See Appendix E for input and training details. We use accuracy as the metric. Since there are equal numbers of positive and negative samples, the accuracy is equivalent to the commonly used balanced accuracy (Luo et al., 2023), which is the average of the accuracy on positive and negative samples. We also report the accuracy of positive and negative samples separately.

3.2 Results

Table 2 reveals significant differences in performance between LLMs tested under zero-shot conditions and smaller models fine-tuned with parameter-efficient methods. Among the zero-shot LLMs, GPT-4o achieved the highest overall accuracy, outperforming Qwen2.5-Plus and DeepSeek-v3. However, even GPT-4o struggled with negative samples, achieving only 70.4% accuracy on the dev set and 71.6% on the test set. This limitation highlights a persistent challenge in distinguishing negative cases, which significantly impacts overall accuracy. DeepSeek-v3, while demonstrating near-perfect accuracy on positive samples, performed poorly on negative samples (39.6% dev, 39.4% test), indicating a clear trade-off between the two categories.

In contrast, smaller models fine-tuned with the training set achieved remarkable improvements, particularly in handling negative samples. Llama-3.1-8B stood out as the top performer, achieving 91.4% accuracy on the dev set and 90.6% on the test set, while maintaining a strong balance between positive and negative samples. These results suggest that the training data effectively addressed the challenges posed by negative samples, enabling the fine-tuned models to achieve significantly higher overall accuracy. Overall, the results underscore the effectiveness of fine-tuning in improving model robustness, particularly for negative samples. The dataset's training data appears to play a crucial role in enhancing model performance, as evidenced by the fine-tuned models' ability to achieve high accuracy across both positive and negative samples. These insights suggest that tailored training strategies and targeted fine-tuning can significantly enhance model capabilities, even for smaller models.

4 Conclusion

In this work, we propose the first large-scale Chinese dataset CITECHECK for citation faithfulness detection. To solve the high-cost problem caused by the lack of negative samples when constructing the dataset using strong RAG systems, we propose the method of data augmentation with two-stage manual annotation. This method allows us to construct a dataset with a balanced number of positive and negative samples at a relatively low cost and guarantees the quality of the dataset. We conduct experiments and validate the quality of the dataset in two aspects: (1) the test samples consisting of the original samples are challenging for detection, and (2) the training samples consisting of the original positive samples and the augmented negative samples can be effectively applied for training.

Limitations

The main limitation of the dataset is the availability of only binary judgment labels (positive or negative). We do not manually label which part of the statement in the negative sample is unsupported, nor do we manually label the evidence in the documents that the statement in the positive sample is supported. However, key segments labeling and modifications in the LLM augmentation phase are available, which compensates for the limitation to some extent.

The main limitation of the experiments is the lack of more experiments on other test sets for the model obtained from training to show the generalization performance. This limitation comes from the lack of relevant Chinese datasets. We will continue to track the relevant Chinese datasets proposed and conduct experiments.

Ethics Statement

We comply with the license to use language models for scientific research purposes only. Questions are collected with the permission of the license of open-source datasets or with the consent of the relevant users. The datasets we construct will also be open source for scientific research purposes. We conduct checks to minimize potential risk issues with datasets, including personal privacy concerns and harmful content.

The AI assistant we use in our work is Copilot (for simple code completion).

References

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 17754–17762. AAAI Press.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023a. RARR: researching and revising what language models say, using language

models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 16477–16508. Association for Computational Linguistics.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6465–6488. Association for Computational Linguistics.

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, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jen-

nifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S. Yu, and Zhijiang Guo. 2024a. Towards understanding factual knowledge of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Xuming Hu, Xiaochuan Li, Junzhe Chen, Yinghui Li, Yangning Li, Xiaoguang Li, Yasheng Wang, Qun Liu, Lijie Wen, Philip Yu, and Zhijiang Guo. 2024b. Evaluating robustness of generative search engine on adversarial factoid questions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10650–10671, Bangkok, Thailand. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8460–8478. Association for Computational Linguistics.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Huanshuo Liu, Hao Zhang, Zhijiang Guo, Kuicai Dong, Xiangyang Li, Yi Quan Lee, Cong Zhang, and Yong Liu. 2024. Ctrla: Adaptive retrieval-augmented generation via probe-guided control. *CoRR*, abs/2405.18727.

Nelson F. Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10,*

2023, pages 7001–7025. Association for Computational Linguistics.

Wen Luo, Tianshu Shen, Wei Li, Guangyue Peng, Richeng Xuan, Houfeng Wang, and Xi Yang. 2024. Halludial: A large-scale benchmark for automatic dialogue-level hallucination evaluation. *CoRR*, abs/2406.07070.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for abstractive text summarization. *CoRR*, abs/2303.15621.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1906–1919. Association for Computational Linguistics.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *CoRR*, abs/2112.09332.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Yujia Qin, Zihan Cai, Dian Jin, Lan Yan, Shihao Liang, Kunlun Zhu, Yankai Lin, Xu Han, Ning Ding, Huadong Wang, Ruobing Xie, Fanchao Qi, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2023. Webcpm: Interactive web search for chinese long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto,*

Canada, July 9-14, 2023, pages 8968–8988. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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

Haochen Tan, Zhijiang Guo, Zhan Shi, Lu Xu, Zhili Liu, Yunlong Feng, Xiaoguang Li, Yasheng Wang, Lifeng Shang, Qun Liu, and Linqi Song. 2024. Proxyqa: An alternative framework for evaluating long-form text generation with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 6806–6827. Association for Computational Linguistics.

Rui Wang. 2023. Dataset: Zhihu-kol.

Bright Xu. 2019. Nlp chinese corpus: Large scale chinese corpus for nlp.

Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 4615–4635. Association for Computational Linguistics.

A TrickQA Details

Real-world questions do not always have the correct premises. For example, in the question "水俣病的传染途径是什么？(What is the route of infection for Minamata disease?)", Minamata disease is not an infectious disease. Taking this situation into account, we add a small number of human-written questions with incorrect premises and LLM-generated questions with hard-to-verify premises in the question collection phase. The number of these questions in the total number of questions is about 3%.

B Supplementary Statistics

B.1 Question Sources

The percentages of different question sources are: WebText (29.5%), WebCPM (33.6%), Zhihu-KOL (29.8%), RGB (3.8%), TrickQA (3.3%).

B.2 The Number of Documents

The percentages of the number of cited documents in one sample are: 1 (74.9%), 2 (15.0%), 3 (5.2%), 4 (2.9%), 5 (2.0%).

C Augmentation Details

See Table 5 for the prompt for LLM augmentation. Table 6 provides an English version.

D Annotation Details

D.1 Instructions for the First Phase

In the first stage, we provide the annotators with the question, answer, statement, and cited documents. What LLM considers to be key segments are highlighted in red in the cited documents. We instruct the annotators to follow the process below:

(1) First look at the highlighted text. If the highlighted text fully supports the statement, then the annotation is positive; if the highlighted text contradicts the statement, then the annotation is negative.

(2) If the annotation cannot be derived from the highlighted text, then look at the rest of the documents to make the annotation. When the documents fully support the statement, the label is positive, and when there is any information in the statement that contradicts the documents or information that is not mentioned in the documents, the label is negative.

D.2 Instructions for the Second Phase

In the second stage, we provide the annotator with the statement and the modified documents. In the documents, the modified parts are highlighted in green, where the dashed and crossed-out text is deleted and the rest is added.

For the annotation of whether the quality of the modification is acceptable, the annotators are instructed to note that qualified modifications need to satisfy the following two requirements: (1) There are no contradictions within each modified document. (2) The modified key segments are fluent in their own right and in the context of the document. The annotation for support is the same as the first stage, but based on the modified documents.

D.3 Label Consistency

Our annotation is carried out in batches. At the end of each batch, we carry out consistency rate checking and provide feedback to the annotators to gradually achieve a labeling consistency rate of over 80%. Samples with inconsistent labeling are discarded to ensure label consistency.

E Input and Training Details

We input the statement and the cited documents into the model and ask the model to determine whether the statement is fully supported by the documents, outputting yes or no. For input, we label and concatenate the cited documents in order (as shown in Table 1). For training, we use the following settings: For training, we use the following settings: learning rate is 5e-4, number of epochs is 10, scheduler is cosine scheduler, warmup ratio is 0.03, batch size is 256, and LoRA setting is $r = 8$, $a = 32$ and 0.1 dropout. We report the model performance for the epoch that achieves the best performance on the dev set.

F Supplementary Experiments

F.1 Shortcut Features

By prompt guidelines in the data augmentation phase and revalidation in the annotation phase, we attempt to prevent shortcut features in the augmented samples that do not reflect the true support relationship. To verify the effectiveness of the control, we supplement the experiment by providing only the statement or documents when fine-tuning LLama-3.1-8B. Table 3 shows the results. The performances when only the statement or the documents are provided are much lower than when

Dev	Acc	Acc_p	Acc_n
statement only	62.2	70.0	54.4
documents only	58.8	59.0	58.6
full sample	91.4	91.6	91.2
Test	**Acc**	**Acc_p**	**Acc_n**
statement only	60.2	65.8	54.6
documents only	58.7	57.4	60.2
full sample	90.6	90.4	90.8

Table 3: Results of experiments on LLama-3.1-8B when providing only the statement or documents for fine-tuning. We show a comparison with using the full sample.

Dev	Acc	Acc_p	Acc_n
GPT-4o (0-shot)	**83.7**	97.0	70.4
GPT-4o (10-shot)	82.4	96.6	68.2
Qwen2.5-Plus (0-shot)	81.6	97.0	66.2
Qwen2.5-Plus (10-shot)	76.7	98.2	55.2
DeepSeek-v3 (0-shot)	69.4	99.2	39.6
DeepSeek-v3 (10-shot)	80.0	98.8	61.2
Test	**Acc**	**Acc_p**	**Acc_n**
GPT-4o (0-shot)	**83.9**	96.2	71.6
GPT-4o (10-shot)	82.1	96.0	68.2
Qwen2.5-Plus (0-shot)	81.2	94.8	67.6
Qwen2.5-Plus (10-shot)	74.4	95.4	53.4
DeepSeek-v3 (0-shot)	69.4	99.4	39.4
DeepSeek-v3 (10-shot)	78.5	98.2	58.8

Table 4: Results of experiments on LLMs Results of experiments on LLMs in 0-shot and 10-shot settings.

the full sample is provided, suggesting that relying only on shortcut features (rather than real support relationships) is insufficient to discriminate the positivity or negativity of the sample well.

F.2 Few-shot Experiments

We supplement the 10-shot (5 positive, 5 negative) experiments on LLMs. Table 4 shows the results. For Qwen-Plus and GPT-4o, the 10-shot setting does not improve the overall accuracy. For DeepSeek-v3, few-shot improves overall accuracy, but the accuracy does not exceed GPT-4o under the 0-shot setting. This complements the support for the conclusions about the difficulty of the samples.

G Related Works

Language models are known to produce hallucinations - statements that are inaccurate or unfounded (Maynez et al., 2020; Hu et al., 2024a). To address this limitation, recent research has focused on augmenting LLMs with external tools such as retrievers (Guu et al., 2020; Borgeaud et al., 2022; Liu et al., 2024) and search engines (Nakano et al., 2021; Komeili et al., 2022; Tan et al., 2024). While this approach suggests that generated content is supported by external references, the reliability of such attribution requires careful examination. Recent studies have investigated the validity of these attributions. Liu et al. (2023) conducted human evaluations to assess the verifiability of responses from generative search engines. Hu et al. (2024b) further investigate the reliability of such attributions when giving adversarial questions to RAG systems. Their findings revealed frequent occurrences of unsupported statements and inaccurate citations, highlighting the need for rigorous attribution verification (Rashkin et al., 2023). However, human evaluation processes are resource-intensive and time-consuming. To overcome these limitations, existing efforts (Gao et al., 2023a,b; Luo et al., 2024) proposed an automated approach using Natural Language Inference models to evaluate attribution accuracy. While several English-language benchmarks have been developed for this purpose (Yue et al., 2023), comparable resources in Chinese are notably lacking. Creating such datasets presents unique challenges, particularly in generating realistic negative samples (unsupported citations). To address this gap, we introduce the first large-scale Chinese dataset for citation faithfulness detection, developed through a cost-effective two-stage manual annotation process.

这里有一段陈述和对应的一段参考文本。请按如下步骤完成任务，严格按我给出的格式进行输出：
（1）找到参考文本中所有直接支撑陈述中信息的原始关键文段（可能有多个，每一处都要找到）。每行输出一个原始关键文段及其直接支撑的陈述中的信息，格式为“关键文段编号：关键文段（支撑陈述中的信息：支撑信息）”。
（2）请将关键文段分组，每组包含的关键文段支撑陈述中的相同或相关的信息，输出一行分组结果，格式为“关键文段分组：第一组：（第一组关键文段编号），第二组：（第二组关键文段编号）...”。例如，陈述中有2个信息，关键文段1支撑信息1，关键文段2支撑信息2，关键文段3支撑信息1，那么输出“关键文段分组：第一组：（1，3），第二组：（2）”
（3）选择一组关键文段，对其中支撑陈述中信息的部分进行修改，满足以下要求：
- 修改应该使得关键文段无法完全支撑陈述中的对应信息。
- 修改应该保持关键文段的逻辑通顺、关键文段中的信息之间不矛盾。
- 修改之后的关键文段应该在参考文本的上下文语境中保持逻辑通顺，且与参考文本中的其他内容不矛盾。
- 只修改支撑陈述中信息的部分，其它部分保持不变。
- 如果一组中有多个关键文段，修改后它们的信息应该保持一致。
你需要尝试两种修改方法：
- 改变信息：将关键文段中的某一处信息修改为另外的信息。请不要进行与原信息产生直接冲突的修改。例如，原信息为“奥迪A7旗舰版的最高速度比上一代快”，合适的修改是“奥迪A7豪华版的最高速度比上一代快”，不合适的修改1是“奥迪A7旗舰版的最高速度比上一代慢”（使用反义词，与原信息直接冲突），不合适的修改2是“奥迪A7旗舰版的最高速度不比上一代快”（添加否定词，与原信息直接冲突）。
- 删除信息：将关键文段中的某一处信息删除。关键文段如果是完整的句子，删除信息后应该仍然是一个完整的句子。例如，原文段为“由于天气原因，项目推迟至3月15日启动”（完整的句子），合适的修改是“由于天气原因，项目推迟至3月启动”（仍然是完整的句子），不合适的修改是“由于天气原因”（不再是完整的句子）。
对每种方法，输出被修改的关键文段，并检查其逻辑通顺程度，给出一个1~10以内的整数作为评分（越高表示越通顺）。每行输出一个修改后的关键文段，格式为“方法-修改后的关键文段编号：修改后的关键文段（逻辑通顺程度：分数）”。

Table 5: The complete prompt for the LLM augmentation.

Here is a statement and a corresponding piece of reference text. Please complete the task as follows, strictly following the format I have given for the output:
(1) Find all the original key passages in the reference text that directly support the information in the statement (there may be more than one, find each one). Output one original key passage per line and the information in the statement it directly supports in the format "Key passage number: key passage (information in the supporting statement: supporting information)".
(2) Please group key passages, each group contains key passages supporting the same or related information in the statement, output one line of the grouping results in the format of "Key passage grouping: Group 1: (first group of key passage numbers), Group 2: (second group of key passage numbers) ...". For example, if there are 2 pieces of information in the statement, key paragraph 1 supports information 1, key paragraph 2 supports information 2, and key paragraph 3 supports information 1, then the output is "Key Paragraph Grouping: Group 1: (1, 3), Group 2: (2)".
(3) Select a group of key text segments and modify the parts of them that support the information in the statement to meet the following requirements:
- The modification should make it impossible for the key passage to fully support the corresponding information in the statement.
- The modifications should maintain the logical flow of the key passages and no contradictions between the information in the key passages.
- The modification should keep the key paragraph logically coherent in the context of the reference text and not contradict the rest of the reference text.
- Modify only the parts that support the information in a statement, leaving the rest unchanged.
- If there is more than one key passage in a set, the information in them should remain consistent after revision.
You need to try two methods of modification:
- Changing the message: modifying the message in one part of the key paragraph to another. Do not make changes that directly conflict with the original information. For example, if the original message is "The Audi A7 Signature Edition has a faster top speed than its predecessor", an appropriate change would be "The Audi A7 Luxury Edition has a faster top speed than its predecessor", and an inappropriate change1 would be "The Audi A7 Signature Edition has a slower top speed than its predecessor" (using an antonym, which is in direct conflict with the original message), and inappropriate modification 2 is 'The top speed of the Audi A7 Signature Edition is not faster than the previous generation' (adding a negative word, which is in direct conflict with the original message).
- Delete Information: Remove information from a place in a key paragraph. If the key paragraph is a complete sentence, it should still be a complete sentence after deleting the information. For example, if the original paragraph reads "Due to weather conditions, the project was delayed until March 15" (complete sentence), an appropriate change would be "Due to weather conditions, the project was delayed until March" (still a complete sentence), an inappropriate change would be "Due to the weather" (no longer a complete sentence).
For each method, output the key passage that was modified and check its logical fluency, giving an integer within 1 to 10 as a rating (higher means more fluent). Output one modified key passage per line in the format "method-modified key passage number: modified key passage (logical fluency: score)".

Table 6: The complete prompt for the LLM augmentation (translated into English).