Do We Need Language-Specific Fact-Checking Models? The Case of Chinese

Anonymous ACL submission

Abstract

This paper investigates the potential benefits of language-specific fact-checking models, focusing on the case of Chinese. We first demonstrate the limitations of translation-based methods and multilingual large language models (e.g., GPT-4), highlighting the need for language-specific systems. We further propose a Chinese fact-checking system that can better retrieve evidence from a document by incorporating context information. To better analyze token-level biases in different systems, we construct an adversarial dataset based on the CHEF dataset, where each instance has large word overlap with the original one but holds the opposite veracity label. Experimental results on the CHEF dataset and our adversarial dataset show that our proposed method outperforms translation-based methods and multilingual LLMs and is more robust toward biases, while there is still large room for improvement, emphasizing the importance of language-specific fact-checking systems[1].

1 Introduction

There has been a growing interest in automated fact-checking in recent years (Graves, 2018; Nakov et al., 2021). While misinformation exists in various languages, the majority of studies have predominantly focused on claims and evidence in English (Guo et al., 2022; Mubashara et al., 2023). Current research in multilingual fact-checking often lacks grounding in real-world claims (Chang et al., 2023) or is constrained to a single domain, like COVID-19-related misinformation (Shahi and Nandini, 2020). Although the X-Fact dataset (Gupta and Srikumar, 2021) encompasses real-world claims in 25 languages, it does not provide verified evidence documents, which are crucial for substantiating the veracity of these claims.

In this paper, we raise the question: *Should we develop language-specific fact-checking models, or can we effectively utilize existing English models by translating claims and evidence into English?* We present a case study focused on Mandarin Chinese to investigate it for two reasons. Firstly, Chinese is widely spoken by over a billion people and possesses unique linguistic characteristics different from English (Yang et al., 2017; Fei, 2023). Secondly, considering the importance of evidence (Borel, 2023), Chinese is the only language other than English that has an evidence-based dataset annotated manually, i.e. CHEF (Hu et al., 2022)).

Original: 广东两名小学生提干，引发大量讨论。 Translated: Two primary school students in Guangdong raised eyebrows (were promoted), sparking discussion. **ChatGPT:** REFUTED **CHEF Label:** SUPPORTED
Claim 1: 中国超八成地下水遭受污染，不能饮用。 (*Over 80% of China's groundwater is polluted and is unfit for drinking.*) Claim 2: 中国高铁辐射严重引发不孕。(*Radiation from China's high-speed rail seriously causes infertility.*) **ChatGPT:** REFUTED **CHEF Label:** SUPPORTED

Table 1: Upper section: the challenge in accurate translation (Red: Incorrect, Blue: Correct); Lower section: the bias of multilingual LLMs towards certain claims.

We first demonstrate the limitations of translation-based methods (e.g. first translating Chinese claims and evidence into English and then applying English fact-checking models on translated data) or multilingual large language models, such as GPT-4 (OpenAI, 2023). Next, we develop a Chinese fact-checking system for CHEF, utilizing a document-level evidence retriever. Our system outperforms state-of-the-art models by 10% in terms of accuracy and Macro F1, and also achieves higher accuracy than using multilingual LLMs. To examine biases in our system, we create an adversarial dataset for Chinese fact-checking. Experiments show a significant decrease in both accuracy and F1 score due to biases often specific to the Chinese culture. Overall, our study highlights the necessity of devising language-specific fact-checking models.

[1] Our dataset and code will be publicly available.

	Retrievers / Verifiers	Semantic Ranker		Document-level Retriever		Gold Evidence	
		Accuracy	Macro F1	Accuracy	Macro F1	Accuracy	Macro F1
Translation	GT+DeBERTa	59.23	59.76	60.15	61.29	66.84	66.57
	GPT-4+DeBERTa	62.17	62.43	62.36	60.01	67.95	67.82
Multilingual LLM	GPT-3.5-Turbo	53.29	51.46	55.45	51.32	58.79	54.97
	GPT-4-Turbo	65.78	62.35	69.17	69.01	73.67	73.96
Chinese Specific	BERT-base	63.00	62.88	67.66	67.66	77.79	77.62
	Attention-based	64.01	63.65	69.00	68.35	78.56	78.46
	Graph-based	62.43	62.42	69.25	69.14	78.95	78.39
	RoBERTa-large	66.37	66.24	72.31	72.31	79.38	79.47
	DeBERTa-large	**69.89**	**68.34**	**74.50**	**74.46**	**81.46**	**81.15**

Table 2: Results on CHEF. For the translated baselines, we first translate the evidence and claims via Google Translator (GT) and GPT-4, then apply the DeBERTa-large claim verifier.

2 Chinese Fact-Checking Systems

To construct a Chinese fact-checking system, two straightforward approaches are direct translation from Chinese to English and the application of multilingual LLMs. However, as demonstrated in Table 1, translation from Chinese to English may result in inaccuracies, particularly with idiomatic expressions or language-specific phrases (Shao et al., 2018). Additionally, LLMs such as ChatGPT, primarily trained on English texts (Lai et al., 2023; Hu et al., 2023), exhibit a bias toward Western perspectives. Table 1 illustrates instances of scientifically refuted claims that GPTs tend to accept, with corresponding retrieved evidence. To examine the abovementioned limitations in a systematic way, we conduct experiments on a large scale Chinese evidence-based dataset, CHEF.

Retrievers We first introduce a novel Document-level retriever to improve the evidence retriever. Unlike previous work that treats evidence selection as pairwise sentence classification in isolation (Hu et al., 2022), we consider the context of the evidence sentences. Inspired by Stammbach (2021), we train a retriever to assign a score to each Chinese token within an evidence document and then aggregate these token scores at the sentence level. In particular, we fine-tune a BigBird (Zaheer et al., 2020) to assign a value of 1 to tokens that belong to annotated evidence for a claim, while assigning a value of 0 to all other tokens. During inference, we compute the average scores for all tokens within each sentence. If the resulting average score exceeds 0.5, we classify the sentence as evidence. We compare our proposed document-level retriever with the Semantic Ranker (Nie et al., 2019; Liu et al., 2020) used by Hu et al. (2022), and utilizes BERT pre-trained on a Chinese corpus.

Verifiers We utilize DeBERTa (He et al., 2021) to verify a claim given the selected evidence, using the Chinese version pretrained on the WuDao Corpora (Wang et al., 2022). We also compare our results with the baselines in Hu et al. (2022), including BERT-base (Devlin et al., 2019), Attention-based (Gupta and Srikumar, 2021), and Graph-based (Liu et al., 2020) methods. We also incorporate the RoBERTa-based model (Liu et al., 2019b), GPT-3.5-Turbo (OpenAI, 2022) and GPT-4-Turbo (OpenAI, 2023) for a more comprehensive comparison. For the GPT models, we use 5 shots for in-context learning. We provide detailed experimental settings in the Appendix A.

Results on CHEF As shown in Table 2, our system that combines Document-level Retriever and DeBERTa-large, yields the best results with an accuracy of 74.50% and a Macro F1 score of 74.46%. There is an improvement of over 10% compared to the best translation-based result (GPT-4+DeBERTa) and 5% over the best multilingual LLM model (GPT-4-Turbo) in both metrics. The results over CHEF emphasize the necessity of language-specific fact-checking tools.

Evidence Retrieval The Document-level Retriever, paired with three different verifiers, improves accuracy and Macro F1 by about 5% over the Semantic Ranker. Regarding the recall of human-annotated gold evidence, Document-level Retriever leads to 10% higher Recall@5 (Table 5). We also find that our new retriever can retrieve evidence pieces which, when considered individually cannot verify the claim but, when combined they can. Table 6 gives a detailed example in the Appendix B.

Claim Verification The pipeline's performance is improved by incorporating RoBERTa and DeBERTa as claim verifiers. The DeBERTa-large yields a notable enhancement, with a 5% uplift in both accuracy and Macro F1 scores over the best-reported baseline with attention-based retriever and document-level verifier.

Word	LMI(10^{-6})	$p(l\|w)$
中国 (China)	1189	0.56
电影 (Movie)	1008	0.84
国际 (International)	629	0.80
发布 (Release/Announce)	599	0.74
金融 (Finance)	593	0.66
亿元 (Hundred Million Yuan)	500	0.66
外交 (Diplomacy/Foreign Affairs)	496	0.85
外交部 (Ministry of Foreign Affairs)	481	0.92
人民币 (RMB/Chinese Yuan)	469	0.84
银行 (Bank)	469	0.63

Word	LMI(10^{-6})	$p(l\|w)$
病毒 (Virus)	1105	0.66
疫苗 (Vaccine)	1013	0.64
台湾 (Taiwan)	962	0.77
可以 (Can/Be able to)	901	0.72
出现 (Appear)	478	0.74
肺炎 (Pneumonia)	475	0.70
手机 (Mobile phone)	451	0.77
冠状 (Coronary)	414	0.93
日本 (Japan)	402	0.72
感染 (Infection)	395	0.66

Table 3: Top 10 LMI-ranked phrases in the train set of CHEF for SUPPORTED (left) and REFUTED (right).

3 Biases in CHEF

To explore the reasons behind the deficiency of translation services and multilingual LLMs, we investigate the biases present in the CHEF dataset in this section. Prior research has demonstrated that fact-checking datasets, such as FEVER (Thorne et al., 2018) and MultiFC (Augenstein et al., 2019), result in training models that rely on heuristics such as surface-level patterns within claims, potentially impeding their ability to generalize effectively (Schuster et al., 2019; Thorne et al., 2019). In this section, we show that while the biases are present as in English language datasets and models, they are specific to the Chinese culture.

First, in CHEF, claims are categorized into domains such as politics, society, health, and culture and we find a significant skew in the distribution: 64% of social and 66% of health claims are REFUTED, while 55% in politics and 72% in culture are SUPPORTED. Notably, there is an imbalance in the proportion of social and health claims, which collectively constitute 68% of the total, compared to the other 3 categories. Figure 1 in the Appendix details the label distribution across domains.

We further examine the correlation between phrases within the claims and the corresponding labels. The word distribution within the training set is analyzed for this purpose. Initially, all claims in the training set are tokenized by Chinese text segmentation tool, jieba[2]. The average length of the words is 2.39 characters. Then, two metrics are employed to assess the correlation between phrases and labels. Following Schuster et al. (2019), first we use $p(l|w)$ to calculate the probability of a label l given the presence of a specific phrase w in the claim. As this metric tends to exhibit bias towards low-frequency words, the second metric utilizes Local Mutual Information (LMI; Evert 2005) to identify high-frequency n-grams that display a strong correlation with a particular label. The $p(l|w)$ and LMI between phrase w and label l is defined as follows:

$$p(l \mid w) = \frac{\text{count}(w, l)}{\text{count}(w)} \quad (1)$$

$$LMI(w, l) = p(w, l) \cdot \log\left(\frac{p(l \mid w)}{p(l)}\right) \quad (2)$$

where we follow Schuster et al. (2019) to estimate $p(l)$ by $\frac{\text{count}(l)}{|D|}$, $p(w, l)$ by $\frac{\text{count}(w,l)}{|D|}$ and $|D|$ is the number of occurrences of all n-grams.

Table 3 lists the top 10 LMI-ranked phrases in the train set of CHEF for SUPPORTED and REFUTED. Prior studies in English datasets, such as Constraint (Patwa et al., 2020), have demonstrated a strong correlation between politician names (*e.g.* Barack Obama and Donald Trump) and refuted claims, however, our research identifies a distinct cultural bias within CHEF. In CHEF, claims about biomedical and health issues frequently exhibit a strong association with negative labels. Terms such as 病毒 (virus), 疫苗 (vaccine), 致癌 (carcinogenic) and 冠状病毒 (coronavirus) are more commonly encountered in refuted claims. Conversely, financial terms like 金融 (finance), 人民币 (RMB), and 央行 (People's Bank of China), as well as political terms such as 中国 (China), 外交部 (Ministry of Foreign Affairs), tend to carry positive labels.

One possible reason behind this is that fact-checking in China tends to avoid criticism of hard-core public issues, such as politics, economics, and other current affairs (Liu and Zhou, 2022). On the contrary, it focuses more on providing references for everyday decision-making, such as in health. Another political reason could be that the Cyberspace Administration of China keeps a close watch on online news services (Liu and Zhou, 2022). Non-state enterprises are not permitted to criticize politics, economics, and other current affairs. Private companies are only authorized to distribute and curate news produced by

[2]https://github.com/fxsjy/jieba

	Original 250 pairs		Generated 750 pairs		Full 1000 pairs	
	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
BERT-base	76.35	75.36	38.56	37.62	49.06	48.72
Attention-based model	78.96	78.12	39.98	39.62	51.01	49.65
Graph-based model	79.55	76.97	39.61	38.67	49.59	49.43
GPT-3.5-Turbo	80.00	55.25	53.73	36.78	60.30	41.39
GPT-4-Turbo	85.60	60.70	**65.20**	47.12	**70.30**	50.73
DeBERTa-large	**86.25**	**85.78**	55.01	**53.03**	62.45	**62.25**

Table 4: Performance comparison of models on the adversarial dataset. The "original 250 pairs" refers to pairs directly extracted from CHEF, while "generated 750 pairs" denotes pairs generated using GPT-4.

state-owned media. Furthermore, in CHEF, certain regions such as 台湾 (Taiwan), 日本 (Japan), and 美国 (United States) are commonly associated with the REFUTED label. This may also reflect the contentious nature of international relations within the realm of Chinese fact-checking.

4 Adversarial Dataset Construction

Our analysis revealed the presence of labels and cultural biases specific to the Chinese context (§ 3). We therefore introduce an adversarial dataset derived from the CHEF dataset for a better evaluation of the models. Inspired by Schuster et al. (2019) and Schuster et al. (2021), to create it we pair each claim-evidence instance with a synthetic counterpart where claim and evidence have high word overlap with the original ones but the opposite veracity label (Figure 2). Under this setting, determining veracity from the claim alone would be equivalent to a random guess. Instead of involving human annotators, we opt for the utilization of GPT-4 to generate the dataset. To control the quality, we invited two Chinese native speakers to annotate randomly sampled 25% of claim-evidence pairs with SUPPORTED, REFUTED or NOT ENOUGH INFO. The results demonstrated strong agreement between humans and GPT-4. They agreed with the dataset labels in 89% of cases, with a Cohen κ of 0.80 (Cohen, 1960). Our approach overcomes labor-intensive manual annotation and rigid rule-based generation, advocating for automated sample generation using LLMs. This new test set nullifies the benefit of relying exclusively on cues from claims. Details of the dataset construction and the prompt we use can be found in Appendix D.

5 Experiments on Adversarial CHEF

Results on Adversarial CHEF Table 4 compares model performance on adversarial versus original data from CHEF. All models perform worse on adversarial examples compared to the original CHEF. Specifically, DeBERTa-large drops from 86.25% accuracy on original pairs to 55.01% and 62.45% on adversarial subsets. Baselines similarly see over 37% decreases in both accuracy and F1 scores. This underscores the models' reliance on surface features and reveals label and cultural biases. Experiments reveal better robustness of GPTs against adversarial datasets. This resilience may stem from GPT models not being fine-tuned on CHEF, thereby avoiding reliance on dataset biases for claim verification. Instead, these models depend more on analyzing retrieved evidence to verify claims. We suggest future research assess systems using both original and our adversarial CHEF dataset for a comprehensive evaluation.

DeBERTa vs. Baselines DeBERTa's performance declines less than that of the baselines including BERT, Attention, and Graph-based models when faced with adversarial examples, about 30% compared to over 37%, suggesting a higher sensitivity to evidence changes. To investigate the reasons behind the decrease in the model's performance, we employ the inoculation fine-tuning method (Liu et al., 2019a). The performance decline observed in the baselines primarily stems from inherent weaknesses within the model family. In contrast, for the DeBERTa model, gradually exposing it to more adversarial samples leads to a gradual reduction in the performance gap. Inoculation results by fine-tuning the model with different sizes of adversarial examples are provided in Figure 3 in the Appendix F.

6 Conclusion

Our study reveals the shortcomings of English-centric fact-checking systems when applied to Chinese claims, highlighting the failure of translation-based methods due to linguistic and cultural nuances. We introduce a novel system that achieves best-reported results on CHEF and provides an adversarial dataset for continued research, underscoring the need for specialized fact-checking models.

Limitations

The performance of our document-level retriever, although enhanced compared to the semantic ranker, is still characterized by a relatively low recall rate. This highlights the persisting challenges in evidence retrieval that require further attention and refinement. Another limitation of our study is the availability of evidence-based fact-checking datasets. We could only conduct our analysis on English- and Chinese-language datasets due to the limited availability of evidence-based datasets in other languages. Consequently, more experiments should be conducted to demonstrate the general applicability of our conclusions.

Ethics Statement

The CHEF dataset employed in our research is accessible to the scientific community, and its use in our experiments presents no conflict of interest. Although, the adversarial dataset used in this study was developed with a GPT-4 model, to ensure its integrity and safety, we conducted an extensive manual review to eliminate sensitive or potentially harmful information. This review received approval from our institution's ethics committee. Furthermore, the hourly salary for annotators surpassed the national minimum wage, and all annotators consented to the use of the data.

References

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.

Brooke Borel. 2023. *The Chicago Guide to Fact-Checking, Second Edition*. University of Chicago Press, Chicago.

Yi-Chen Chang, Canasai Kruengkrai, and Junichi Yamagishi. 2023. XFEVER: Exploring fact verification across languages. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 1–11.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bosheng Ding, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq Joty, and Boyang Li. 2022. Is gpt-3 a good data annotator? *arXiv preprint arXiv:2212.10450*.

Stefan Evert. 2005. The statistics of word cooccurrences: word pairs and collocations.

Yunxi Fei. 2023. The differences in thinking patterns between english and chinese in chinese students' english writing. In *2nd International Conference on Education, Language and Art (ICELA 2022)*, pages 277–285. Atlantis Press.

Isa Fulford and Andrew Ng. 2023. Chatgpt prompt engineering for developers. *DeepLearningAI Blog*.

Lucas Graves. 2018. Understanding the promise and limits of automated fact-checking. *Reuters Institute for the Study of Journalism*.

Zhijiang Guo, Michael Sejr Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Trans. Assoc. Comput. Linguistics*, 10:178–206.

Ashim Gupta and Vivek Srikumar. 2021. X-fact: A new benchmark dataset for multilingual fact checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 675–682, Online. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S. Yu, and Zhijiang Guo. 2023. Do large language models know about facts? *ArXiv*, abs/2310.05177.

Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and Philip Yu. 2022. CHEF: A pilot Chinese dataset for evidence-based fact-checking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3362–3376, Seattle, United States. Association for Computational Linguistics.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, pages 294–297.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Nelson F. Liu, Roy Schwartz, and Noah A. Smith. 2019a. Inoculation by fine-tuning: A method for analyzing challenge datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2171–2179, Minneapolis, Minnesota. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

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

Yusi Liu and Ruiming Zhou. 2022. "let's check it seriously": Localizing fact-checking practice in china. *International Journal of Communication*, 16(0).

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351, Online. Association for Computational Linguistics.

Akhtar Mubashara, Schlichtkrull Michael, Guo Zhijiang, Cocarascu Oana, Simperl Elena, and Vlachos Andreas. 2023. Multimodal automated fact-checking: A survey. *arXiv preprint arXiv:2305.13507*.

Preslav Nakov, David P. A. Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. *CoRR*, abs/2103.07769.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6859–6866.

OpenAI. 2022. Chatgpt blog post.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint*.

Parth Patwa, Shivam Sharma, Srinivas PYKL, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2020. Fighting an infodemic: Covid-19 fake news dataset.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. AVeritec: A dataset for real-world claim verification with evidence from the web. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3419–3425, Hong Kong, China. Association for Computational Linguistics.

Gautam Kishore Shahi and Durgesh Nandini. 2020. FakeCovid – a multilingual cross-domain fact check news dataset for covid-19. In *Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media*.

Yutong Shao, Rico Sennrich, Bonnie Webber, and Federico Fancellu. 2018. Evaluating machine translation performance on Chinese idioms with a blacklist method. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Dominik Stammbach. 2021. Evidence selection as a token-level prediction task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 14–20, Dominican Republic. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2019. The FEVER2.0 shared task. In *Proceedings of the*

Second Workshop on Fact Extraction and VERification (FEVER), pages 1–6, Hong Kong, China. Association for Computational Linguistics.

Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, Chongpei Chen, Ruyi Gan, and Jiaxing Zhang. 2022. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? gpt-3 can help. *arXiv preprint arXiv:2108.13487*.

Man Yang, North Cooc, and Li Sheng. 2017. An investigation of cross-linguistic transfer between chinese and english: a meta-analysis. *Asian-Pacific Journal of Second and Foreign Language Education*, 2:1–21.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297.

A Experiment Setup

In the results presented in Table 2, the translation models initially employ Google/GPT-4 to convert all claims and evidence within the CHEF dataset to English. Subsequently, an English RoBERTa-large is fine-tuned to assess the veracity of these claims using the CHEF training set. For multilingual LLMs, we apply a five-shot in-context learning approach with both GPT-3.5-Turbo and GPT-4-Turbo. Regarding the baseline models—BERT-base, attention-based, and graph-based models—we adhere to the default hyperparameters as delineated in the CHEF study (Hu et al., 2022). We run our experiments on A100-SXM-80GB GPUs. For each pipeline system, we conduct three independent experiments and report the mean values.

B Comparison of Different Retrievers

Table 5 compares the performance of Semantic Ranker and Document-level Retriever. The Document-level Retriever leads to better *Recall@5* and *Marco F1*. *Recall@5* measures the proportion of gold evidence that are successfully retrieved among the top 5 retrieved evidence sentences.

Although outperforming the Semantic Ranker, the Document-level Retriever only attains a 33.58% *Recall*@5, indicating the difficulty of evidence retrieval, yet remarkably leads to a 74.46% Macro F1 score in claim verification. This may be due to the CHEF's gold evidence annotation not being exhaustive, a known issue in datasets with evidence retrieved from the Web (Schlichtkrull et al., 2023), and thus the retriever can return correct evidence that was not annotated. Additionally, the model might leverage surface-level patterns in claims to inform verification, which allows for high accuracy even when the available evidence is insufficient.

Sentence Retrieval	Recall@5	Macro F1
Semantic Ranker	21.24 $\pm$2.13	70.58 $\pm$1.56
Document-level Retriever	**33.58** $\pm$**2.08**	**74.46** $\pm$**1.78**

Table 5: Comparison of Semantic Ranker and Document-level Retriever for evidence sentence retrieval with DeBERTa-large.

Table 6 is an example where leveraging document-level information can help with the evidence retrieval. To verify the claim: "运用红酒含有花青素的原理，可以简单检测红酒的真假。(*The principle that red wine contains anthocyanins allows for a straightforward authenticity test.*)", each retriever collects five pieces of evidence. Without additional context, it is not possible to retrieve the sentences highlighted in red through semantic matching alone. None of these sentences, when considered individually, can be used to verify the claim. However, when taken together, they provide a comprehensive explanation of why anthocyanins can be utilized to test red wine. Having access to the entire document makes it much easier to accurately predict similar examples.

C Generative AI in Annotation Tasks

Generative AI models, such as ChatGPT [3], DELL-E (Ramesh et al., 2021), have witnessed significant advancements in recent years, enabling the generation of high-quality content across various modalities, including text, speech, video, and images. Notably, OpenAI's release of GPT-4 (OpenAI, 2023) has demonstrated human-level performance on diverse professional and academic benchmarks.

Given the remarkable ability to generate new content based on human instructions, researchers have explored the potential of employing generative AI models as a substitute for labour-intensive annotation tasks. For instance, Huang et al. (2023)

[3] https://chat.openai.com/

Semantic Ranker	Document-level Retriever
有一个妙招，一秒钟鉴定红酒真假 (*There's a clever trick, one-second wine authenticity test*)	而假红葡萄酒中多由酒精、糖精和香精色素勾兑而成，里面不含花青素 (*Fake red wine is often made by blending alcohol, glycerin, and artificial colorants, without containing anthocyanins*)
这时，如果红酒变成深蓝色，就是真红酒；如果没有反应，则是假红酒 (*At this point, if the red wine turns deep blue, it's genuine; if there's no reaction, it's fake*)	由于真正的红葡萄酒中含有丰富的花青素 (*Because authentic red wine contains abundant anthocyanins*)
如何辨别真假红酒，教你简单一招 (*How to distinguish real from fake red wine, teaching you a simple trick*)	花青素在酸性条件下呈现紫红色，而在碱性条件下呈现蓝绿色 (*Anthocyanins appear purplish-red under acidic conditions and bluish-green under alkaline conditions*)
若是色素勾兑的红酒，颜色则无变化 (*If it's red wine adulterated with colorants, the color remains unchanged*)	其实，还有一个更简单的方法没说 (*In fact, there's an even simpler method not mentioned*)
把用水兑开的食用碱水滴在红酒上面； (*Drip food-grade alkali water diluted with water onto the red wine;*)	如果我们家里的红酒用食用碱检测没有变色，那么基本可以肯定你买到了假酒 (*If our home red wine doesn't change color when tested with food-grade alkali, then it's safe to say you've bought fake wine*)

Table 6: Evidence sentences retrieved by Semantic Ranker and Document-leverl Retriever for the claim: "运用红酒含有花青素的原理，可以简单检测红酒的真假(*The principle that red wine contains anthocyanins allows for a straightforward authenticity test.*)."

examined the use of ChatGPT in providing natural language explanations (NLEs) for detecting implicit hateful speech. Their findings reveal that ChatGPT accurately identifies 80% of implicit hateful tweets, and in cases of disagreement, the experimental results indicate a higher alignment between ChatGPT's outputs and lay people's perceptions. Moreover, Wang et al. (2021) and Ding et al. (2022) highlight the feasibility and cost-effectiveness of leveraging generative AI models for data labelling tasks. Their research emphasizes the potential benefits of incorporating these models into the annotation workflow.

To generate high-quality content, the selection of an appropriate prompt is crucial. A prompt refers to a set of instructions provided to a Large Language Model (LLM) to customize, enhance, or refine its capabilities (Liu et al., 2023). In our task, the prompt essentially represents how we provide instructions to the GPT-4 models. Different prompts can significantly impact the model's performance (Liu et al., 2023). Kojima et al. (2022) have even demonstrated that simply adding the phrase "Let's think step by step" before each answer can enhance the quality of the generated content.

D Adversarial Dataset Construction

D.1 Task Definition

To further detect and eliminate bias in CHEF, we propose to generate a new Chinese adversarial dataset for it. We adopt the methodology presented by (Schuster et al., 2019) as our primary framework for constructing a symmetrical dataset for CHEF, as illustrated in Figure 2. Our approach involves generating synthetic claim-evidence pairs that maintain the same relationship (e.g., SUPPORTS or REFUTES) while conveying contrasting factual information. Moreover, we ensure that each sentence in the new pair exhibits the inverse relationship with its corresponding sentence in the original pair.

Some new rules have been devised to better suit the Chinese context. More specifically, when rewriting the given claim "陈大文在北京称，2020年版第五套人民币5元纸币将发行，防伪性能提升。" (Chen Dawen, announced that the 2020 edition of the fifth series of 5-yuan banknotes will be issued, with improved anti-counterfeiting features, in Beijing.), in our framework, the following rewriting strategies are allowed:

- Important nouns that appear in both the claim and the evidence can be modified. These include key information such as time, place, person, and number. Changing these essential

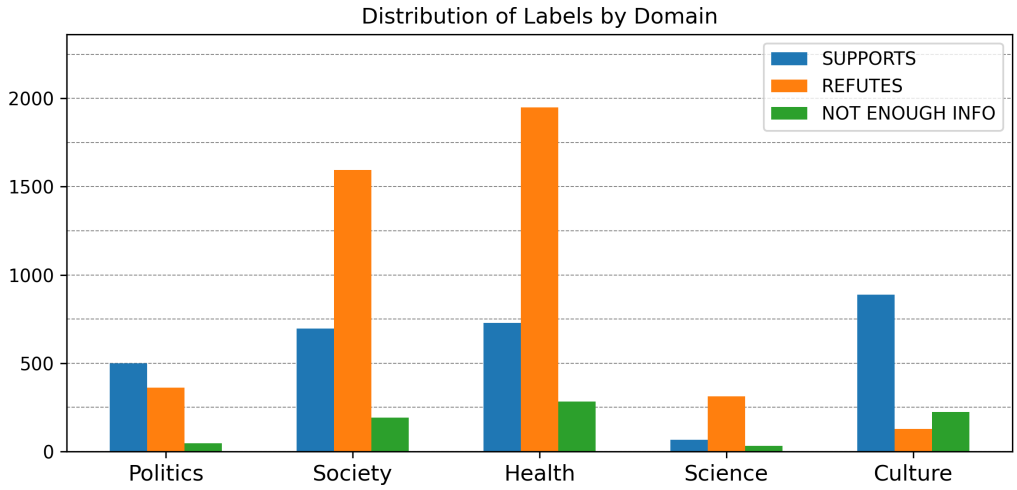

Figure 1: The distribution of labels across different domains in CHEF.

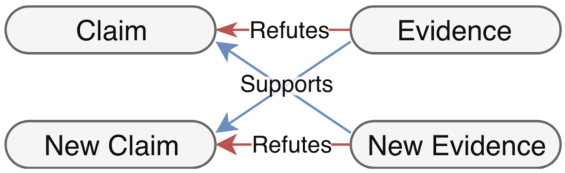

Figure 2: A illustration of the relationship between the original pair and the generated pair (Schuster et al., 2019).

terms can alter the original meaning of the sentence. For example, substituting the name "Chen Dawen" with "Li Xiaoming," revising the year "2020" to "2023," replacing the location "Beijing" with "Shanghai," and transforming the denomination "5 yuan" to "10 yuan."

- Verbs or phrases indicating degrees in both the claim and the evidence can be replaced with their opposites. For instance, substituting "rise" with "fall," changing "increase" to "decrease," converting "helpful" to "unhelpful," replacing "substantiated" with "unsubstantiated," and transforming "no evidence" to "evidence not found."

Note that these methods do not constitute an exhaustive set of legal rewrite methods. They serve as heuristics for the model, which may also employ similar modifications automatically. Similarly, the evidence undergoes a comparable rewriting process. For additional examples of these methods, please refer to Table 7. To rewrite the sentences, we employ the state-of-the-art GPT-4 (OpenAI, 2023) model ,which has demonstrated human-level performance in various NLP tasks. By leveraging the GPT-4 model, we eliminate the laborious task of human annotation and enhance the diversity of generation through handcrafted rules.

E Prompt Engineering

Given the importance of prompt engineering for the quality of the generated data, as well as the scarcity of relevant literature, it is imperative to carefully craft our prompt. To address this challenge, we sought guidance from the empirical findings of the open source community [4], which provided valuable insights into prompt design practices. Furthermore, we consult the recently published prompt design guideline by (Fulford and Ng, 2023) to ensure our approach aligns with the newest recommendations. We conducted extensive experiments to iteratively refine our prompt, culminating in the development of an innovative prompt that not only enhances the quality of generated results but also exhibits versatility, enabling its seamless adaptation to a wide range of tasks.

According to Fulford and Ng (2023), the effectiveness of a prompt relies on two key principles. **Principle 1** emphasizes the significance of providing clear and specific instructions to the model. To achieve this, the prompt should employ delimiters (such as backticks) to clearly demarcate distinct parts of the input. Furthermore, the provision of examples helps the model formulate a "few-shot" prompt, allowing it to generate responses based on limited examples. **Principle 2** focuses on opti-

[4] https://github.com/f/awesome-chatgpt-prompts

Source	Claim	Evidence	Label
ORIGINAL	2021年12月31日人民币对美元汇率中间价上调27个基点。*On December 31, 2021, the central parity rate of the Chinese yuan against the US dollar was increased by 27 basis points.*	新华社上海12月31日电来自中国外汇交易中心的数据显示，31日人民币对美元汇率中间价报6.5782，较前一交易日上调27个基点。*Shanghai, December 31st (Xinhua) - Data from the China Foreign Exchange Trading System showed that the central parity rate of the Chinese yuan against the US dollar was set at 6.5782 on the 31st, representing an increase of 27 basis points compared to the previous trading day.*	SUPPORT
GENERATED	2021年12月31日人民币对美元汇率中间价下调27个基点。*On December 31, 2021, the central parity rate of the Chinese yuan against the US dollar was decreased by 27 basis points.*	新华社上海12月31日电来自中国外汇交易中心的数据显示，31日人民币对美元汇率中间价报6.5782，较前一交易日下调27个基点。*Shanghai, December 31st (Xinhua) - Data from the China Foreign Exchange Trading System showed that the central parity rate of the Chinese yuan against the US dollar was set at 6.5782 on the 31st, representing a decrease of 27 basis points compared to the previous trading day.*	SUPPORT
ORIGINAL	奥密克戎对抗体中和作用不存在逃逸现象。*There is no evidence of escape phenomenon in the neutralizing action of omicron antibodies.*	结果发现，奥密克戎变异株能被实验中所有单克隆抗体有效中和，没有出现逃逸现象。*The findings revealed that the omicron variant can be effectively neutralized by all monoclonal antibodies tested in the experiment, with no observed escape phenomenon.*	SUPPORT
GENERATED	奥密克戎对抗体中和作用存在大量逃逸现象。*There is a significant amount of escape phenomenon in the neutralizing action of omicron antibodies.*	结果发现，奥密克戎变异株能完全抵抗或部分抵抗实验中所有单克隆抗体的中和作用。*The results indicate that the omicron variant can completely or partially resist the neutralizing action of all monoclonal antibodies tested in the experiment.*	SUPPORT
ORIGINAL	2020年4月，某男子在公园挖土被警方罚款200元。*In April 2020, a man was fined 200 yuan by the police for digging soil in the park.*	经讯问，邓某承认该微博所述情节均为伪造，其本人并未到过绿博园，更没有被公安机关处罚。*After questioning, Mr Deng admitted that the Weibo post was fabricated, and he had never been to Green Park nor been penalized by the police.*	REFUTE
GENERATED	2020年4月，某男子在公园挖土被警方制止，但并未罚款。*In April 2020, a man was stopped by the police for digging soil in the park but was not fined.*	据警方透露，某男子于2020年4月在公园非法挖土，发现后被警方罚款200元。*According to the police, the man was found engaging in unauthorized soil excavation in the park in April 2020 and was subsequently fined 200 yuan.*	REFUTE

Table 7: Examples from the symmetric adversarial dataset are provided to illustrate claim-evidence pairs where the relationship described in the right column is maintained. By combining the generated sentences with the original ones, two additional cases are formed, each with labels that are opposite to one another. The red texts in Chinese highlight the differences between the claim/evidence before and after the rewrite.

mizing the model's processing by breaking down the full task into several subtasks. This approach guides the model to think step by step, enhancing its performance. The structure of our prompt is outlined in Table 8.

E.1 Quality Control

Following the data generation process, we generated 250 new claim and evidence pairs. By permuting them under the symmetric setting Schuster et al. (2019), we obtained an adversarial dataset consisting of 1000 pairs. We then enlisted the participation of two Chinese native speakers to perform annotations on a randomly selected subset of 300 claim-evidence pairs removing their labels, which accounted for 30% of the total pairs within the symmetric adversarial dataset. These annotations involved assigning one of two labels, namely SUPPORTS, and REFUTES, while also flagging instances of nongrammatical cases. The average agreement between the annotators and the pre-existing dataset labels reached 89% of the cases, resulting in a Cohen κ coefficient of 0.80 (Cohen, 1960). It demonstrates that the new claim-evidence pairs generated by GPT-4 mostly remain in their original relation, proving the effectiveness of our method. Additionally, approximately 4% of the cases exhibited minor grammatical errors or typos.

E.2 Error Analysis

After manually examining the wrongly predicted cases for the DeBERTa-large model following the inoculation process, we have identified three primary challenges that current models struggle to address:

- Subtle modifications can induce a dramatic change in sentence meaning. In the adversarial CHEF dataset, a large number of state-

Explanation of Prompt Design	Prompt Snippet
Introduce the background of the task and the input format of the data. Define a role for the model.	我希望你作为一个编辑部的事实核查记者，完成以下的数据标注任务，同时改写声明和证据，使得其各自的含义与原意相反...（I would like you, as an fact-checking journalist, to complete the following annotation task: rewriting claims and evidence so that their respective meanings are the opposite of what they originally meant...）
Give the requirement of how to rewrite the claim.	第一步：修改声明内容，使得其变成于之前含义相反的内容...(Step 1: Modify the claim to make it have the opposite meaning as before...)
Give the requirement of how to rewrite the evidence accordingly.	第二步：对应修改后的声明，修改证据的内容...(Step 2: Modify the evidence accordingly, corresponding to the modified claim...)
Give a detailed example and possible rewrite strategies.	针对例句：[例子]，以下我提供几个理想且合法的修改示例: ...(For the exemplary sentence: [EXAMPLE], I offer the following examples of ideal and legal modifications: ...)
Give a small bunch of human-annotated samples.	请同时参考以下一些其他例句：示例一；示例二；示例三；...(Please also refer to the following additional example sentences: Example 1; Example 2; Example 3; ...)
Emphasize the key requirement.	你可以使用上述例子中的修改方式，也可以使用其他修改方法。但是最重要的是要求修改后的证据仍然能支持修改后的声明。(You can use the modification strategies mentioned above as well as other ways to make the changes. However, the most important aspect is to ensure that the modified evidence still supports the modified claim.)
Give the claim and evidence pair that is needed to rewrite delimited by triple backticks.	``` TEXT ```

Table 8: This table outlines the purpose of each snippet in the prompt, explaining the role of each section according to the prompt design principles.

ments exhibit slight differences before and after modifications, often differing by only one or two Chinese characters. Given the rich semantic nature of Chinese characters, even a single-word alteration can reverse the entire sentence's meaning. For instance, in the first example of Table 9 and the first example of Table 7, minor changes involving a single character completely alter the original meaning. These nuanced distinctions pose difficulties for models to accurately capture. Furthermore, even if these changes are encoded in the model's parameters, they may not receive significant weighting during veracity assessment.

- Adversarial CHEF includes numerical reasoning challenges that lack a dedicated mechanism. While the original CHEF dataset contains extensive instances of numbers, there are relatively few statements that necessitate inference from numerical information. In contrast, the adversarial CHEF dataset introduces numerous modifications associated with numbers, requiring the model to determine whether the statements align with the evidence's numerical values. For example, consider the second example in Table 9. However, our current approaches lack a dedicated mechanism to address these numerical issues, resulting in numbers being treated similarly to text.

- Inferences from implicit or circumstantial evidence present challenges in assessing the claims. In most cases, the evidence is straightforward, enabling easy judgment of the statement's correctness. However, there are instances where the evidence used for inference does not explicitly provide the truth of the statement or directly contradict its content. For instance, the third example in Table 9 does not directly specify what is incorrect with the statement (e.g., mentioning that it should be 50,000 instead of 70,000). Instead, the evidence uses terms like "non-representative" and "sensationalized" to indirectly point out the unreasonableness of the data results. It is important to note the distinction between this type of challenge and cases involving "not

Source	Claim	Evidence	Label
ORIGINAL	2019年1月，成都万象城车祸致一人死亡。*In January 2019, a car accident at Chengdu The MixC Mall resulted in one fatality.*	经交警分局反馈：核实现场无人死亡，只有一个伤者。*According to feedback from the Traffic Police, upon verification, there were no fatalities at the scene, with only one injured individual.*	REFUTE
GENERATED	2019年1月，成都万象城车祸无人死亡。*In January 2019, the car accident at Chengdu The MixC Mall resulted in no fatalities.*	经交警分局反馈：核实现场一人死亡，还有一个伤者。*According to feedback from the Traffic Police, upon verification, there was one fatality at the scene, as well as one injured individual.*	REFUTE
ORIGINAL	2018年春节档总票房累计20.36亿，“就地过年”让影院更火爆。*During the 2018 Spring Festival season, the total box office revenue reached 2.036 billion RMB, making the cinemas even more popular with the "celebrate the Lunar New Year locally" trend.*	票房方面，2018年春节档，中国电影票房20.36亿，打破2017年春节档创下的15.06亿票房纪录，创春节档票房新纪录。*In terms of box office performance, the 2018 Spring Festival season achieved a record-breaking box office revenue of 2.036 billion RMB, surpassing the previous record of 1.506 billion RMB set in the 2017 Spring Festival season and establishing a new record for the Spring Festival box office.*	SUPPORT
GENERATED	2021年春节档总票房累计78.45亿，“就地过年”让影院更火爆。*During the 2021 Spring Festival season, the total box office revenue reached 7.845 billion RMB, making the cinemas even more popular with the "celebrate the Lunar New Year locally" trend.*	票房方面，2021年春节档，中国电影票房78.45亿，打破2019年春节档创下的59.06亿票房纪录，创春节档票房新纪录。*In terms of box office performance, the 2021 Spring Festival season achieved a record-breaking box office revenue of 7.845 billion RMB, surpassing the previous record of 5.906 billion RMB set in the 2019 Spring Festival season and establishing a new record for the Spring Festival box office.*	SUPPORT
ORIGINAL	2021年全国有七万硕士在送外卖。*In 2021, there were 70,000 individuals with master's degrees working as food delivery drivers nationwide.*	就这样，两个并不具有代表性的“1%”，被自媒体简单渲染成“全国七万硕士在送外卖”。*Just like that, two non-representative “1%" were sensationalized by the media as “70,000 master's degree holders nationwide working as food delivery drivers."*	REFUTE
GENERATED	2021年全国有七万硕士在送外卖为谣言。*The claim that there were 70,000 master's degree holders working as food delivery drivers nationwide in 2021 is a rumour.*	就这样，两个并不具有代表性的“1%”，得出了一个较为科学的估算，即“全国七万硕士在送外卖”。*Thus, two non-representative “1%" have led to a more scientific estimate of “70,000 master's degree holders delivering nationwide".*	REFUTE

Table 9: Wrongly predicted cases of the DeBERTa-large model after the inoculation process. The red texts in Chinese highlight the differences between the claim/evidence before and after the rewrite.

enough information," where the former can be deduced through careful inference. Effectively addressing this type of problem requires models with stronger reasoning capabilities.

F Inoculation by fine-tuning

Upon evaluating the model with synthetic datasets, it's clear the model underperforms compared to the original benchmarks. The precise weaknesses that these datasets reveal are not immediately revealed. To understand this better, we adopt the method of inoculation by fine-tuning, introduced by Liu et al. (2019a). This method allows models to be exposed to a small portion of challenging dataset data to see how the performance changes.

Post-inoculation, we anticipate three possible outcomes:

Outcome 1: A narrowing of the performance discrepancy between the original and challenge test sets suggests that the challenge data didn't expose model weaknesses but rather a lack of diversity in the original dataset.

Outcome 2: No change in performance on either test set indicates that the challenge dataset has pinpointed a fundamental model flaw, as the model fails to adjust even when familiarized with the challenge data.

Outcome 3: A performance drop on the original test set suggests the fine-tuning skewed the model to suit the challenge data, highlighting a deviation from the original data characteristics. This could be due to differences in label distribution or annotation artifacts that are dataset-specific.

Figure 3 shows results from fine-tuning with various amounts of adversarial data. We observe the "performance gap" as the difference in model performance on the original versus adversarial test sets pre-inoculation.

For BERT-base, attention-based, and graph-based models, we observe minor performance changes—Outcome 2—signifying that fine-tuning does not close the performance gap significantly,

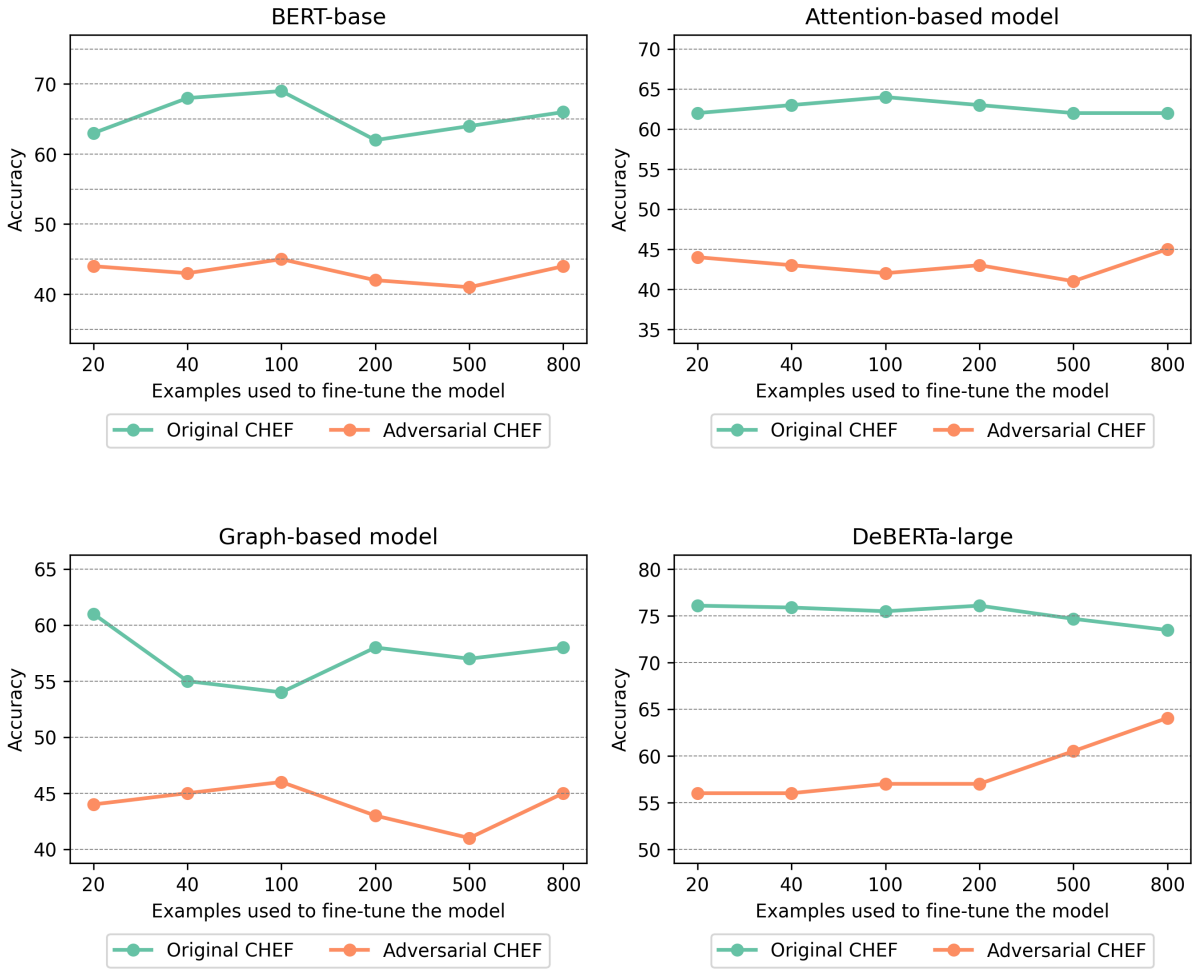

Figure 3: Inoculation results by fine-tuning the model with different sizes of adversarial examples. To evaluate the models, we employ both the original CHEF test set and the adversarial CHEF test set.

pointing to a core weakness in adapting to adversarial data distributions.

In contrast, the DeBERTa-large model shows a reduced performance gap post-inoculation, cutting it down by 53% after fine-tuning with 800 adversarial examples. Its strong performance on the original dataset persists, suggesting DeBERTa's architecture, with its nuanced attention to content, relative, and absolute positions in sentences, equips it to handle slight alterations in claim or evidence more adeptly.