MEDFACT: A Large-scale Chinese Dataset for Evidence-based Medical Fact-checking of LLM Responses

Tong Chen[1,3,*], **Zimu Wang**[2,3,*], **Yiyi Miao**[1,3], **Haoran Luo**[1], **Yuanfei Sun**[1], **Wei Wang**[2], **Zhengyong Jiang**[1,†], **Procheta Sen**[3,†], **Jionglong Su**[1,†]
[1]School of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University, China
[2]School of Advanced Technology, Xi'an Jiaotong-Liverpool University, China
[3]Department of Computer Science, University of Liverpool, United Kingdom
{Tong.Chen19,Zimu.Wang19}@student.xjtlu.edu.cn
{Zhengyong.Jiang02,Jionglong.Su}@xjtlu.edu.cn, Procheta.Sen@liverpool.ac.uk

Abstract

Medical fact-checking has become increasingly critical as more individuals seek medical information online. However, existing datasets predominantly focus on human-generated content, leaving the verification of content generated by large language models (LLMs) relatively unexplored. To address this gap, we introduce MEDFACT, the first evidence-based Chinese medical fact-checking dataset of LLM-generated medical content. It consists of 1,321 questions and 7,409 claims, mirroring the complexities of real-world medical scenarios. We conduct comprehensive experiments in both in-context learning (ICL) and fine-tuning settings, showcasing the capability and challenges of current LLMs on this task, accompanied by an in-depth error analysis to point out key directions for future research. Our dataset is publicly available at https://github.com/AshleyChenNLP/MedFact.

1 Introduction

Nowadays, over one-third of American adults have sought medical information online before consulting a healthcare professional (Fox and Duggan, 2012). However, the intentional proliferation of medical misinformation presents substantial risks to public health. As a result, medical fact-checking has emerged as a critical task to verify the authenticity of online medical content. This process involves assessing both the medical claims and the supportive or refuted evidence to enhance transparency in medical information and mitigate the spread of misinformation (Zhao et al., 2024).

Existing medical fact-checking datasets have primarily focused on human-generated content. With the advent of large language models (LLMs) and their growing use in medical counseling (Wang et al., 2024a, 2025; Na et al., 2025), these datasets are not equipped to address the distinct challenges posed by LLM-generated content, where the embedded parametric knowledge often lacks clear evidence or precise medical details, resulting in the hallucination issue and factual inaccuracies (Peng et al., 2023; Li et al., 2024b). Moreover, these datasets remain limited in scale, making it challenging to effectively train and evaluate LLMs for medical fact-checking tasks, with only 300 and 750 samples in CoVERT (Mohr et al., 2022) and HealthFC (Vladika et al., 2024), respectively. Overall, the limited adaptability and scale of existing datasets restrict their practicality, highlighting a significant research gap in this area.

To tackle the aforementioned challenges, we introduce MEDFACT, the first evidence-based Chinese medical fact-checking dataset designed for medical content generated by LLMs. As shown in Figure 1, we begin by collecting medical questions from the webMedQA dataset (He et al., 2019), for which we generate responses using LLMs. These responses are subsequently decomposed and decontextualized to isolate individual claims. We then assess the check-worthiness of each claim and retrieve relevant evidence for verification with an "LLM-then-Human" approach, ensuring both efficiency and quality. Based on the dataset, we conduct extensive experiments in both in-context learning (ICL, Brown et al., 2020) and fine-tuning settings, demonstrating the effectiveness of large-scale models in leveraging parametric knowledge and the adaptability of smaller models for this task.

Our main contributions are as follows: (1) We introduce MEDFACT, the first evidence-based Chinese dataset targeting LLM-generated medical content. (2) We conduct extensive experiments to showcase existing LLMs on this task and highlight the challenges by reasoning-oriented models. (3) We present a thorough error analysis to identify key areas for future research, including handling medical ambiguity, recognizing semantic containment,

[*]Equal contribution.
[†]Corresponding authors.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 32329–32342
November 4-9, 2025

and understanding medical synonymy.

2 Related Work

Medical Fact-checking. Existing medical fact-checking datasets primarily focus on content generated by humans. SciFact (Wadden et al., 2020) reformulates expert-written claims in biomedical literature and pairs them with summaries that provide evidence. HEALTHVER (Sarrouti et al., 2021) specializes in public health, particularly for verifying claims concerning health advice and treatments. CoVERT (Mohr et al., 2022) targets the verification of medical claims during COVID-19, identifying and validating misinformation during this global health crisis. HealthFC (Vladika et al., 2024) is a bilingual dataset focused on health-related claims, annotated by medical experts and supported by systematic reviews and clinical trials. However, fact-checking datasets based on LLM-generated content in the medical domain remain unexplored, posing a significant research gap in the era of LLMs.

Fact-checking of LLM Responses. With the increasing prevalence of LLMs, recent research has shifted to evaluate the factuality of LLM-generated context. HaluEval (Li et al., 2023) is designed to evaluate factuality around three tasks: knowledge-based discourse, summarization, and world knowledge question answering. Attributable to Identified Sources (AIS, Rashkin et al., 2023) emphasizes the factuality of dialogue systems with pre-injected background knowledge. Under the open-domain setting, FELM (Chen et al., 2024) centers long-form responses with fine-grained factuality annotations. BingCheck (Li et al., 2024a) utilizes human annotations within the SELF-CHECKER framework. Factcheck-Bench (Wang et al., 2024b) further spans three levels of granularity. In this paper, we inherit the emphasis on fact-checking LLM responses, but our position in the medical field remains underexplored in existing research.

3 Dataset Construction

Figure 1 presents the overall pipeline for constructing our MEDFACT dataset, consisting of six distinct steps. In this section, we introduce each of the steps in detail.

3.1 Problem Definition

The evidence-based medical fact-checking task is defined as the following: Given a set of claims $C = \{c_1, c_2, \ldots, c_n\}$ and their evidence $E = \{e_1, e_2, \ldots, e_n\}$, where c_i is the i-th claim and e_i is its corresponding evidence, our goal is to learn a function $f : C \times E \rightarrow \mathcal{L}$, where $\mathcal{L} = \{$SUPPORTED, PARTIALLY SUPPORTED, REFUTED, UNCERTAIN, NOT APPLICABLE$\}$. For each pair (c_i, e_i), models should predict the correct label (i.e., veracity) $y_i = f(c_i, e_i) \in \mathcal{L}$, thereby determining the degree to which the evidence supports or refutes the claim.

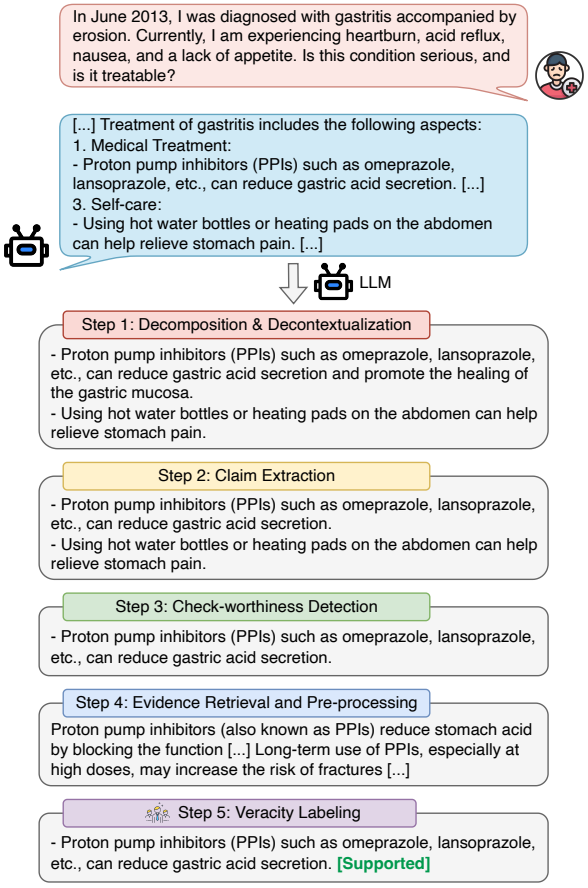

Figure 1: Overall pipeline for constructing the MEDFACT dataset.

3.2 Dataset Construction

Data Collection and LLM Responses Generation. We begin by sourcing biomedical questions from an existing dataset, webMedQA (He et al., 2019), from which we randomly select a subset of $1,500$ questions. These questions span a broad range of 23 medical topics, such as internal medicine, surgery, ophthalmology, and mental health, ensuring diversity in the represented medical information. For each question $q_i \in Q$, we leverage Yi (`Yi-Large-Turbo`, Young et al., 2025), an LLM with performance comparable to GPT-4 but significantly more cost-effective to generate responses $r_i = \text{LLM}(q_i)$, following prior work (Kang et al., 2024). Here, Q denotes the collection

of selected questions and r_i signifies the generated response corresponding to q_i.

Decomposition and Decontextualization. Once the questions and their responses are obtained, for each response r_i, we utilize DeepSeek-V2.5 (Liu et al., 2024a), a model with enhanced language understanding and instruction-following capabilities, to decompose the response into a sequence of discrete statements $\{s_1, s_2, \ldots, s_p\}$, where p denotes the total number of individual statements. This process ensures that each statement is self-contained, without any irrelevant background information or extended context, thereby eliminating potential interference in subsequent analysis.

Claim Extraction. To extract claims for verification, we reframe claim identification as a generation task rather than a classification task. Specifically, given a group of statements $\{s_1, \ldots, s_p\}$ derived from the previous step of "Decomposition and Decontextualization," we employ DeepSeek-V2.5 to generate a set of claims $\{c_1, \ldots, c_m\}$, where each c_i corresponds to a declarative statement that needs to be verified based on external evidence. Here, $m \leq p$, since each generated claim is derived from a specific source statement, but not all source statements necessarily yield a claim. This approach leverages the generation capabilities of LLMs to extract declarative content, in line with prior work (Chern et al., 2023; Li et al., 2024a). We also conduct a human evaluation on a subset of 100 questions, focusing on the steps of "Decomposition and Decontextualization" and "Claim Extraction," with an accuracy of 100% on these sampled questions, ensuring the high quality of both the dataset and all intermediate steps.

Check-worthiness Detection. Not all claims warrant verification, as some may be self-evident or trivial, while others are more significant or contentious. Therefore, we introduce a claim check-worthiness detection mechanism to identify claims that merit further verification, involving evaluating each claim c_i based on four key factors: *Popularity*, *Public Interest*, *Impact*, and *Timeliness* (Husain et al., 2020). The check-worthiness of a claim is framed as a binary classification task, where DeepSeek-V2.5 is prompted to assign a binary label to each claim, indicating whether it should proceed to the subsequent stages. To ensure correctness, we manually check and revise the samples according to the aforementioned criteria to determine the final set of check-worthy claims.

Type	Train	Val	Test	Overall
# of Samples	924	199	198	1,321
# of Claims	5,064	1,186	1,159	7,409
Text Length (Words)				
Response (Avg.)	563.14	557.51	569.57	563.25
Claim (Avg.)	23.64	23.23	22.74	23.43
Evidence (Avg.)	448.85	439.71	438.42	445.75
Veracity Distribution (%)				
Supported	66.86	66.95	64.02	66.43
Partially Supported	20.14	21.84	22.17	20.73
Refuted	0.91	0.93	1.29	0.97
Uncertain	10.31	8.94	10.09	10.06
Not Applicable	1.78	1.35	2.42	1.81

Table 1: Statistics of the MEDFACT dataset.

Evidence Retrieval and Pre-processing. Afterward, we retrieve evidence from the web that may either support or refute the claims. For each check-worthy claim c_i, we use the Google Search API[1] to retrieve the top three most relevant documents. To minimize the impact of irrelevant information on the verification process, we leverage GLM-4-Long (Zeng et al., 2024a), renowned for its long-document understanding capabilities, to extract the most pertinent and high-quality evidence sentences and consolidate them into coherent, self-contained evidence e_i for each claim.

Veracity Labeling. In the final stage, we annotate the veracity y_i given a claim c_j and its associated evidence e_i to represent the truthfulness of each claim based on the evidence, where the label is assigned in a pre-defined set including *supported*, *partially supported*, *refuted*, *uncertain*, and *not applicable*, whose definitions are as follows:

- *Supported*: Evidence fully supports the claim.
- *Partially Supported*: Some sentences support the claim with uncertainties.
- *Refuted*: Any evidence contradicts the claim.
- *Uncertain*: Relate to the claim but no sentences refute, support, or partially support it.
- *Not Applicable*: Completely irrelevant.

To bootstrap the annotation process, we propose an "LLM-then-Human" approach, where we first generate preliminary labels using GLM-4-Long. These initial labels are then reviewed and refined by two trained undergraduate student annotators with medical backgrounds via an annotation platform based on Label Studio[2]. We randomly select 200 samples from each annotator's work and have them

[1] https://www.googleapis.com/
[2] https://labelstud.io/

Model	#Para.	Overall Performance				Per-class F1-score				
		Accuracy	Precision	Recall	F1-score	SUP	PAR	REF	UNC	NOT
Human Performance										
Human Performance	–	88.88	74.50	81.06	77.02	95.82	78.75	84.21	69.65	56.67
In-context LLMs										
GPT-4o	–	67.53	**46.18**	**52.12**	**46.47**	82.57	45.07	45.45	**40.22**	19.05
GPT-4o mini	–	70.97	43.95	45.65	42.79	**84.25**	45.22	39.02	33.33	12.12
Qwen3-30B-A3B	30B	54.77	41.58	48.31	40.26	67.56	52.34	48.48	3.51	**29.41**
Qwen3-32B	32B	57.71	43.20	47.55	41.15	70.27	**55.63**	**51.61**	4.00	24.24
GLM-4-32B	32B	70.01	48.45	46.37	46.81	83.05	45.42	42.42	46.51	16.67
Deepseek-V2.5	236B	60.54	50.40	37.97	32.03	76.43	48.79	11.11	3.85	20.00
DeepSeek-V3	671B	**67.66**	46.05	39.27	37.37	83.21	47.69	25.00	14.06	16.90
Fine-tuned LLMs										
Qwen2.5-7B	7B	68.51	51.75	41.08	43.52	81.77	**47.00**	40.00	48.83	0.00
Meditron3-Qwen2.5-7B	7B	68.77	47.67	39.50	41.44	82.42	46.43	30.00	48.37	0.00
Qwen3-4B	4B	68.51	56.16	41.11	44.65	81.85	46.74	38.10	44.79	**11.76**
Qwen3-8B	8B	65.66	50.20	38.80	41.31	79.72	43.11	30.00	48.00	5.71
InternLM3-8B	8B	67.73	48.87	**41.69**	**44.03**	81.93	46.23	**41.67**	40.82	9.52
GLM-4-9B	9B	**69.20**	**57.13**	39.30	42.43	**82.61**	45.42	33.33	**50.76**	0.00

Table 2: Experimental results of in-context and fine-tuned LLMs on the MEDFACT dataset, in which the best performance on each type of LLM is highlighted in **bold**. (SUP: Supported; PAR: Partially Supported; REF: Refuted; UNC: Uncertain; NOT: Not Applicable)

reviewed by a third annotator. The Inter-Annotator Agreement, measured by Cohen's Kappa (Cohen, 1960), is 81.54%. More complete examples for the construction process are illustrated in Appendix C.

3.3 Dataset Analysis

As detailed in Table 1, the MEDFACT dataset consists of 1,321 medical questions[3] and 7,409 claims. The dataset is carefully partitioned into a split of 70%:15%:15% to ensure independence across subsets. The average text length of medical responses, claims, and evidence is 563.14, 23.64, and 448.85 words, respectively. For the veracities, the supported claims dominate (66.43%), followed by partially supported (20.73%), then uncertain, applicable, and refuted, reflecting the subtle nature of medical misinformation and posing significant challenges when dealing with this uneven distribution, mirroring the complexities of real-world medical scenarios. Examples of the dataset are depicted in Appendix A.

4 Experiments

4.1 Experimental Setup

We conduct experiments in in-context learning (ICL, Brown et al., 2020) and fine-tuning settings, with the prompt detailed in Figure 4. Under the ICL setting, we evaluate GPT-4o (`2024-08-06`, Hurst et al., 2024), GPT-4o mini (`2024-07-18`), Qwen3-30B-A3B (Yang et al., 2025), Qwen3-32B, GLM-4-32B (`0414`, Zeng et al., 2024b), DeepSeek-V2.5 (Liu et al., 2024a), and DeepSeek-V3 (`0324`, Liu et al., 2024b), where the temperature is set as 0.7 for all models. For fine-tuning, we focus on Qwen2.5-7B, Meditron3-Qwen2.5-7B[4], Qwen3-4B, Qwen3-8B, InternLM3 (Cai et al., 2024), and GLM-4-9B (`0414`). During fine-tuning, we set the number of epochs as 10, the learning rate as $1e-4$, the batch size as 4, and the number of gradient accumulation steps as 4. Reasoning is disabled for Qwen3 in all experiments. Performance is assessed using accuracy, macro precision, macro recall, and macro F1-score. All experiments are conducted on 2 NVIDIA A800 Tensor Core GPUs. Prompts for the experiments are organized in Appendix B.

4.2 Experimental Results

Table 2 illustrates the results of the experimented LLMs on the MEDFACT dataset. We also sample a subset of 2,000 claims to verify the human performance on this task. From the table, we have the following observations:

(1) GLM-4-32B demonstrates the best performance in the ICL setting, and Qwen3-4B achieves optimal results among the fine-tuned models. However, none of the models outperform humans, underscoring the significant challenge posed and the

[3]The rest of the 179 questions cannot be successfully processed because of the content moderation policy of LLMs (He et al., 2024).

[4]https://huggingface.co/OpenMeditron/Meditron3-Qwen2.5-7B

necessity of introducing the MEDFACT dataset. Despite the significant disparity in model sizes (671B for DeepSeek-V3 and 4B for Qwen3), smaller models outperform larger ones after fine-tuning. This highlights the effectiveness of large-scale models in leveraging parametric knowledge for medical fact-checking, as well as the adaptability of smaller models to this specific task. Meditron3-Qwen2.5, specialized in clinical medicine, does not outperform Qwen2.5, which may be attributed to its improved capabilities in medical question answering, potentially at the expense of other tasks.

(2) LLMs perform effectively in classifying "supported," followed by "partial supported" and "refuted" evidence, while the "uncertain" and "not applicable" veracities exhibit the poorest performance. After fine-tuning, the Qwen3, InternLM3, and GLM-4 models show performance gains in classifying "uncertain" evidence, though this comes at the cost of other evidence, highlighting a significant challenge in achieving simultaneous improvements across all evidence categories.

4.3 Error Analysis

To point out promising avenues for future research, we conduct a thorough analysis of errors made by LLMs and categorize them into three types:

(1) **Evidence Misunderstanding:** LLMs often misclassify "uncertain" instances into other veracity categories. As shown in Table 3, the evidence discusses the function of AST but omits the normal range, which should be classified as "uncertain," but LLMs incorrectly assign "partially supported" to such cases, highlighting their drawbacks in handling ambiguous or incomplete evidence.

(2) **Semantic Containment Overlook:** LLMs fail to recognize the semantic containments between claims and evidence. As shown in Table 4, the evidence exclusively supports meditation as a method for stress relief, but does not address the broader claim, resulting in the incorrect assignment of "partially supported," rather than the appropriate "supported" classification.

(3) **Medical Synonymy Misjudgment:** LLMs often struggle to identify semantic equivalence between the formal medical terms and their commonly used aliases. As shown in Table 5, despite the evidence clearly describing the cyclical nature of the disease, the model incorrectly predicts a label of "uncertain" instead of "supported" because of its failure to interpret the synonymous relationship between the two terms.

Claim	The normal range for aspartate aminotransferase (AST) is typically between 8-40 U/L.
Evidence	Aspartate aminotransferase (AST) is an enzyme that helps the body break down amino acids. [...] This test is sometimes referred to as SGOT (Serum Glutamic-Oxaloacetic Transaminase).
Prediction	Partially Supported
Label	Uncertain

Table 3: Example of Evidence Misunderstanding.

Claim	Daily stress can be managed through methods such as meditation, yoga, and deep breathing exercises.
Evidence	If stress makes you feel anxious, tense, and worried, try meditation. [...] Meditation is most commonly used for relaxation and stress relief. It is considered a beneficial complementary therapy for both the mind and body. Meditation can help you deeply relax and calm your mind. [...] However, meditation should not be used as a substitute for medical treatment.
Prediction	Supported
Label	Partially Supported

Table 4: Example of Semantic Containment Overlook.

Claim	The symptoms of psoriasis may periodically worsen and improve.
Evidence	Psoriasis is a common, chronic (long-term) disease with no known cure. It can be painful, interfere with sleep, and make it difficult to concentrate. [...] Common triggers for individuals with a genetic predisposition to psoriasis include infections, cuts or burns, and certain medications.
Prediction	Uncertain
Label	Supported

Table 5: Example of Medical Synonymy Misjudgment.

5 Conclusion and Future Work

We introduce MEDFACT, the first evidence-based Chinese medical fact-checking dataset for LLM-generated medical content, consisting of $1,321$ questions and $7,409$ claims, mirroring the complexities of real-world medical scenarios. Experimental results in both ICL and fine-tuning settings showcase the capability and challenges of current LLMs on this task, and we perform an in-depth error analysis to point out key directions for future research. In the future, we will propose innovative methodologies to deal with the identified errors.

Limitations

Although MEDFACTS pioneers the research in medical fact-checking of LLM responses, its scope is currently limited to Chinese because the medical questions are sourced from the webMedQA dataset. While this limitation does not diminish our contribution and the validity of our findings, we advocate for further research efforts to develop more diverse datasets with multilinguality. Furthermore, similar to earlier fact-checking datasets, the label distribution of MEDFACT is imbalanced. Future work can focus on generating synthetic data or applying adversarial learning techniques to inject misinformation (Pan et al., 2023; Wang et al., 2024c) to alleviate this limitation.

Ethical Considerations

We discuss the following ethical considerations related to our MEDFACT dataset as follows: (1) **Intellectual Property.** The webMedQA dataset is distributed under the Apache-2.0 license[5], which is free for research use. We follow the regulations of the license and will share our dataset under Apache-2.0 upon publication. (2) **Annotators Treatments.** We hired student annotators and fairly pay them according to agreed salaries and workloads. (3) **Intended Use.** MEDFACT can be utilized to develop more persuasive models in the field of medical fact-checking. Researchers can also inherit our dataset design to develop their own datasets. (4) **Controlling Potential Risks.** Since the documents of MEDFACT do not contain private information and the annotation process is not necessary to make many judgments about social risks, we believe MEDFACT does not introduce any additional risks. We manually verified some randomly sampled data to ensure the dataset did not contain risky issues.

Acknowledgments

We thank all anonymous reviewers for their valuable feedback. This research is funded by the Postgraduate Research Scholarship (PGRS) at Xi'an Jiaotong-Liverpool University, contract number FOSSP221001.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Shiqi Chen, Yiran Zhao, Jinghan Zhang, I-Chun Chern, Siyang Gao, Pengfei Liu, and Junxian He. 2024. Felm: benchmarking factuality evaluation of large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. Factool: Factuality detection in generative ai – a tool augmented framework for multi-task and multi-domain scenarios. *Preprint*, arXiv:2307.13528.

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

Susannah Fox and Maeve Duggan. 2012. Health online 2013. *Pew Research Internet Project Report*.

Jianfei He, Lilin Wang, Jiaying Wang, Zhenyu Liu, Hongbin Na, Zimu Wang, Wei Wang, and Qi Chen. 2024. Guardians of discourse: Evaluating llms on multilingual offensive language detection. *Preprint*, arXiv:2410.15623.

Junqing He, Mingming Fu, and Manshu Tu. 2019. Applying deep matching networks to chinese medical question answering: a study and a dataset. *BMC medical informatics and decision making*, 19:91–100.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Iltifat Husain, Blake Briggs, Cedric Lefebvre, David M Cline, Jason P Stopyra, Mary Claire O'Brien, Ramupriya Vaithi, Scott Gilmore, and Chase Countryman. 2020. Fluctuation of public interest in covid-19 in the united states: Retrospective analysis of google trends search data. *JMIR Public Health Surveill*, 6(3):e19969.

Xiaoqiang Kang, Zimu Wang, Xiaobo Jin, Wei Wang, Kaizhu Huang, and Qiufeng Wang. 2024. Template-driven llm-paraphrased framework for tabular math word problem generation. *Preprint*, arXiv:2412.15594.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language

[5] https://www.apache.org/licenses/LICENSE-2.0

models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Miaoran Li, Baolin Peng, Michel Galley, Jianfeng Gao, and Zhu Zhang. 2024a. Self-checker: Plug-and-play modules for fact-checking with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 163–181, Mexico City, Mexico. Association for Computational Linguistics.

Ruosen Li, Zimu Wang, Son Tran, Lei Xia, and Xinya Du. 2024b. Meqa: A benchmark for multi-hop event-centric question answering with explanations. In *Advances in Neural Information Processing Systems*, volume 37, pages 126835–126862. Curran Associates, Inc.

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. 2024a. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *Preprint*, arXiv:2405.04434.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024b. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Isabelle Mohr, Amelie Wührl, and Roman Klinger. 2022. CoVERT: A corpus of fact-checked biomedical COVID-19 tweets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 244–257, Marseille, France. European Language Resources Association.

Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. A survey of large language models in psychotherapy: Current landscape and future directions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376, Vienna, Austria. Association for Computational Linguistics.

Liangming Pan, Wenhu Chen, Min-Yen Kan, and William Yang Wang. 2023. Attacking open-domain question answering by injecting misinformation. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 525–539, Nusa Dua, Bali. Association for Computational Linguistics.

Hao Peng, Xiaozhi Wang, Jianhui Chen, Weikai Li, Yunjia Qi, Zimu Wang, Zhili Wu, Kaisheng Zeng, Bin Xu, Lei Hou, and Juanzi Li. 2023. When does in-context learning fall short and why? a study on specification-heavy tasks. *Preprint*, arXiv:2311.08993.

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. *Computational Linguistics*, 49(4):777–840.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024. HealthFC: Verifying health claims with evidence-based medical fact-checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8095–8107, Torino, Italia. ELRA and ICCL.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Yuqi Wang, Qiuyi Chen, Haiyang Zhang, Wei Wang, Qiufeng Wang, Yushan Pan, Liangru Xie, Kaizhu Huang, and Anh Nguyen. 2024a. Biomedical information retrieval with positive-unlabeled learning and knowledge graphs. *ACM Trans. Intell. Syst. Technol.*

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. 2024b. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14199–14230, Miami, Florida, USA. Association for Computational Linguistics.

Zimu Wang, Hongbin Na, Rena Gao, Jiayuan Ma, Yining Hua, Ling Chen, and Wei Wang. 2025. From posts to timelines: Modeling mental health dynamics from social media timelines with hybrid LLMs. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 249–255, Albuquerque, New Mexico. Association for Computational Linguistics.

Zimu Wang, Wei Wang, Qi Chen, Qiufeng Wang, and Anh Nguyen. 2024c. Generating valid and natural adversarial examples with large language models. In *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 1716–1721.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao,

Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. 2025. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. 2024a. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. 2024b. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Xiaoyan Zhao, Lingzhi Wang, Zhanghao Wang, Hong Cheng, Rui Zhang, and Kam-Fai Wong. 2024. PACAR: Automated fact-checking with planning and customized action reasoning using large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12564–12573, Torino, Italia. ELRA and ICCL.

A Dataset Examples

Supported Claim Example

Claim: Scurvy is caused by vitamin C deficiency.
Evidence: Scurvy is rare in the United States and may occur in people with alcohol use disorders and in malnourished older adults. Adults with vitamin C deficiency who lack vitamin C in their diets feel easily fatigued, sluggish and irritable, and may experience weight loss, muscle wasting and joint pain. After several months of vitamin C deficiency, scurvy's can develop. A low vitamin C diet that lasts for several months can cause scurvy, which manifests itself as subcutaneous bleeding (especially around hair follicles or the appearance of bruising), bleeding gums, and bleeding in the joints [...]
Veracity: Supported

Partially Supported Claim Example

Claim: Daily stress can be managed through methods such as meditation, yoga, and deep breathing exercises.
Evidence: If stress makes you feel anxious, tense, and worried, try meditation. [...] Meditation is most commonly used for relaxation and stress relief. It is considered a beneficial complementary therapy for both the mind and body. Meditation can help you deeply relax and calm your mind. [...] It allows you to calm the scattered thoughts that crowd your mind and cause stress. This process can improve both mental and physical health. Meditation can help you maintain a peaceful, serene, and tranquil state of mind, which is beneficial for emotional well-being and overall health. [...] However, meditation should not be used as a substitute for medical treatment.
Veracity: Partially Supported

Refuted Claim Example

Claim: There are usually no strict restrictions on sexual activity during the recovery period after an abortion.
Evidence: After an abortion, it is recommended to avoid sexual intercourse for at least one month. During this period, even the use of condoms is not advisable, as the reproductive system requires time to heal. Engaging in sexual activity too soon can lead to infections, bleeding, or even a subsequent pregnancy. [...] After one month, a follow-up examination should be conducted at the hospital to ensure the reproductive system has fully recovered, after which normal sexual activity can be resumed.
Veracity: Refuted

Uncertain Claim Example

Claim: The normal range for aspartate aminotransferase (AST) is typically between 8-40 U/L.
Evidence: Aspartate aminotransferase (AST) is an enzyme that helps the body break down amino acids. Like alanine aminotransferase (ALT), AST is usually present in low levels in the blood. Elevated AST levels may indicate liver damage, liver disease, or muscle injury. This test is sometimes referred to as SGOT (Serum Glutamic-Oxaloacetic Transaminase).
Veracity: Uncertain

Not Applicable Claim Example

Claim: Reducing stress, maintaining a healthy lifestyle, and getting adequate sleep may help improve sexual function.
Evidence: A heart-healthy lifestyle can help prevent cardiac damage that may trigger certain arrhythmias. Reducing and managing stress, controlling high blood pressure, high cholesterol, and diabetes, and maintaining adequate sleep are essential. The recommended sleep goal for adults is 7 to 9 hours per day.
Veracity: Not Applicable

Figure 2: Examples of the MEDFACT dataset with different veracity labels.

B Prompts

B.1 Prompts for Dataset Constuction

Prompt for Generating LLM Responses

```
Question: {given_question} (Please do not need a final summary like 'To summarize', ...)
```

Prompt for Decomposition and Decontextualization

```
Decompose the following text into a sequence of discrete sentences. Each sentence should
be self-contained and clearly express a single piece of information. Remove any irrelevant
background information or extended context. The output should be in one paragraph without
numbering:
<in-context examples>
TEXT: {given_response}
```

Prompt for Claim Extraction

```
A claim is a statement that asserts something as true or false and can be verified with evidence.
Your task is to accurately identify and extract every claim from the following text. Provide
the extracted claim(s) without additional context or irrelevant details. If there are multiple
claims, separate them clearly. Your response MUST be a list of dictionaries. Each dictionary
should contains the key 'claim', which correspond to the extracted claim.
<in-context examples>
TEXT: {processed_response}
```

Prompt for Check-worthiness Detection

```
You are tasked with evaluating whether a claim is 'check-worthy' based on several factors. For
each claim, consider the following:
1. Popularity: The level of circulation or discussion of a claim. Determine whether it is
commonly shared or debated online or in the media.
2. Public Interest: The general public's interest in the outcome or verdict of the claim.
Consider whether people would want to know if the claim is true or false.
3. Impact: The potential impact of verifying or debunking the claim. Evaluate whether it would
influence people's decisions, behaviors, or beliefs.
4. Timeliness: The relevance of a claim to current events, trends, or discussions. Assess
whether its truth or falsehood needs to be determined quickly due to its relation to ongoing
topics.
Given each claim, using above factors to label it as 'Yes' which means check-worthy or 'No'
which means not check-worthy [no need explanation]
<in-context examples>
TEXT: {given_claim}
```

Prompt for Veracity Labeling

```
Please determine the relationship between the following claim and evidence, and assign an
appropriate label: The labels include:
1. Supported: Evidence fully supports
2. Partially Supported: Partial support with uncertainties
3. Refuted: Evidence contradicts
4. Uncertain: Evidence is insufficient or does not clearly indicate the truthfulness of the
claim
5. Not Applicable: Irrelevant evidence
<in-context examples>
Claim: {given_claim}
Evidence: {given_evidence}
Please provide the most appropriate label without giving an explanation.
```

Figure 3: Prompts used for the construction of the MEDFACT dataset.

B.2 Prompt for Experiments

Prompt for ICL and Fine-tuning Experimental Settings

```
You are a professional fact-checking assistant. Given a claim and corresponding evidence,
select the appropriate label from these options:
1. Supported: Evidence fully supports
2. Partially Supported: Partial support with uncertainties
3. Refuted: Evidence contradicts
4. Uncertain: Evidence is insufficient or does not clearly indicate the truthfulness of the
claim
5. Not Applicable: Irrelevant evidence

Respond strictly as:
Judgment: <label>
Claim: {given_claim}
Evidence: {given_evidence}
```

Figure 4: Prompt for in-context learning (ICL) and fine-tuning experiments.

C Dataset Construction Example

This section presents a real case from our MEDFACT dataset. Table 6 displays a question sourced from the webMedQA dataset (He et al., 2019), which we utilize to obtain the original response presented in Table 7. Following the processes of decomposition and decontextualization, we arrive at the processed response shown in Table 8. Tables 9-18 illustrate the final processed results, which encompass claims, corresponding evidence sources, processed evidence, machine label, and human label. These results were generated through a series of steps, including claim extraction, check-worthiness detection, evidence retrieval and pre-processing, and finally, veracity labeling.

What are the common symptoms of liver disease? My father usually maintains strict exercise habits and has always been in good health. He follows a regulated diet and rarely eats out. Recently, he has been experiencing mental fatigue, slight weight loss, and yellowish complexion. We are concerned about potential liver disease.

Table 6: Question of the example from the MEDFACT dataset.

Liver disease can manifest in various symptoms. Here are some common symptoms of liver disease: Jaundice: Liver disease may cause abnormal bilirubin metabolism, leading to yellowing of the skin and sclera. Fatigue: Impaired liver function can result in ineffective energy metabolism, causing persistent tiredness. Anorexia: As the liver participates in digestion, hepatic diseases may affect appetite and digestive functions. Weight Loss: Reduced appetite in liver disease patients often leads to significant weight reduction. Abdominal Pain: Liver disease may cause discomfort or pain in the right upper quadrant. Hepatomegaly: Hepatic enlargement may present as a palpable abdominal mass. Pruritus: Cholestasis caused by liver disease can induce skin itching. Bleeding Tendency: Impaired clotting factor synthesis may lead to gingival bleeding, epistaxis, or subcutaneous hemorrhage. Dark Urine: Urine may appear tea-colored due to liver dysfunction. Pale Stools: Stools may become clay-colored in liver disease. Hepatic Encephalopathy: Severe cases may manifest mental status changes including confusion or coma. Based on your description, your father exhibits symptoms of fatigue, weight loss, and a yellowish complexion, which may indicate liver disease, but could also be signs of other health issues. It is recommended to take your father to the hospital for a comprehensive examination as soon as possible, including liver function tests, blood tests, and ultrasound, to determine the exact cause and receive appropriate treatment.

Table 7: Original response of the example from the MEDFACT dataset.

Jaundice: Liver disease may cause abnormal bilirubin metabolism, leading to yellowing of the skin and sclera.
Fatigue: Impaired liver function can result in ineffective energy metabolism, causing persistent tiredness.
Anorexia: As the liver participates in digestion, hepatic diseases may affect appetite and digestive functions.
Weight Loss: Reduced appetite in liver disease patients often leads to significant weight reduction.
Abdominal Pain: Liver disease may cause discomfort or pain in the right upper quadrant.
Hepatomegaly: Hepatic enlargement may present as a palpable abdominal mass.
Pruritus: Cholestasis caused by liver disease can induce skin itching.
Bleeding Tendency: Impaired clotting factor synthesis may lead to gingival bleeding, epistaxis, or subcutaneous hemorrhage.
Dark Urine: Urine may appear tea-colored due to liver dysfunction.
Pale Stools: Stools may become clay-colored in liver disease.
Hepatic Encephalopathy: Severe cases may manifest mental status changes including confusion or coma.

Table 8: Decomposed and decontextualized response of the example from the MEDFACT dataset.

Claim: *Liver disease may cause darkening of urine resembling tea color.*

Evidence Sources:
- **Source 1**: "Normal urine color is clear to pale yellow. However, certain factors can change the color of urine..."
- **Source 2**: "The whites of eyes and skin typically appear yellow in jaundice patients due to high bilirubin levels..."
- **Source 3**: "Jaundice serves as a warning sign of systemic disease, manifesting as yellow skin discoloration or scleral icterus..."

Processed Evidence: Dark or orange urine may indicate liver dysfunction, particularly when accompanied by pale stools and jaundice.
Machine Label: Partially Supported
Label: Uncertain

Table 9: Claim for the example from the MEDFACT dataset.

Claim: *Liver disease may lead to abnormal bilirubin metabolism and yellowing of the skin and whites of the eyes.*

Evidence Sources:
- **Source 1**: "Jaundice occurs when the liver is diseased and is unable to remove bilirubin in sufficient amounts. Bilirubin is a metabolic waste product from the blood..."
- **Source 2**: "The whites of the eyes and skin usually look yellow in people with jaundice. Jaundice occurs when there is a high level of bilirubin (a yellow pigment) in the blood..."
- **Source 3**: "Jaundice is a warning sign of physical illness. When the skin becomes abnormally yellowish brown or the whites of the eyes turn yellow, this symptom should not be ignored..."

Processed Evidence: Jaundice is an abnormal condition of the body, mainly caused by the increase of bilirubin in the blood...
Machine Label: Supported
Label: Supported

Table 10: Claim for the example from the MEDFACT dataset.

Claim: *Liver disease may present with lighter, off-white colored stools.*

Evidence Sources:
- **Source 1**: "Orange urine may indicate a problem with the liver or bile ducts; look for light-colored stools as well..."
- **Source 2**: "Dark brown or orange urine, yellowish skin and eyes, and whitish stools may indicate liver deficiency..."
- **Source 3**: "If the liver does not produce bile, or if bile is stagnant in the liver, the stools will be light-colored or white..."

Processed Evidence: Dark or orange urine may indicate liver dysfunction, particularly when accompanied by pale stools and jaundice.
Machine Label: Supported
Label: Supported

Table 11: Claim for the example from the MEDFACT dataset.

Claim: *When liver function is impaired, the body may not be able to metabolize energy efficiently, leading to fatigue and lethargy.*

Evidence Sources:
- **Source 1**: "Energy metabolism may be affected, nutrient absorption may deteriorate, and the body may feel more and more tired and sluggish. Changes in urine and feces: When the liver is damaged, it is unable to process wastes and metabolites efficiently, leading to urination..."
- **Source 2**: "If left untreated, over time, hypothyroidism can lead to other health problems, such as high cholesterol and heart problems..."
- **Source 3**: "Fatigue and tiredness caused by liver cancer cannot be eliminated even if the patient lies down and rests for a long time. The main reason for fatigue is that cancer cells damage the metabolism and detoxification function of the liver..."

Processed Evidence: The liver is responsible for storing and releasing energy. When the liver is damaged, energy metabolism may be affected, nutrient absorption becomes poorer, and the body may feel more and more tired and sluggish.
Machine Label: Supported
Label: Supported

Table 12: Claim for the example from the MEDFACT dataset.

Claim: *The liver is involved in the digestive process and liver disease may affect appetite and digestion.*

Evidence Sources:
- **Source 1**: "Statins are highly effective and safe for most patients, but some patients experience drug-related muscle pain, digestive problems, and mental fogginess. In rare cases, liver damage may result..."
- **Source 2**: "The liver is involved in the metabolic processes in the body, and after suffering from hepatitis, the function of bile secretion decreases ... Gastrointestinal dysfunction and other symptoms, which in turn affects the patient's food digestion and absorption..."
- **Source 3**: "The major organs of the digestive system include the liver, stomach, gallbladder, colon and small intestine..."

Processed Evidence: The liver is involved in the metabolic process in the body, after suffering from hepatitis, the function of bile secretion is reduced, which affects the digestion of fat, so there will be anorexia, gastrointestinal dysfunction, etc., which affects the patient's food digestion and absorption.
Machine Label: Supported
Label: Supported

Table 13: Claim for the example from the MEDFACT dataset.

Claim: *People with liver disease may experience loss of appetite, which can lead to weight loss.*

Evidence Sources:
- **Source 1**: "Cirrhosis is the extensive destruction of the internal structure of the liver caused by the permanent replacement of large amounts of normal liver tissue by nonfunctional scar tissue..."
- **Source 2**: "Patients with cirrhosis may experience loss of appetite, weight loss, fatigue and general malaise..."
- **Source 3**: "Obesity increases your risk for diseases that can lead to cirrhosis, such as non-alcoholic fatty liver disease and..."

Processed Evidence: People with cirrhosis may experience symptoms such as loss of appetite, weight loss, fatigue and general malaise. These symptoms may be caused by impaired liver function, as the liver becomes less able to process medications, toxins, and waste products from the body...
Machine Label: Supported
Label: Supported

Table 14: Claim for the example from the MEDFACT dataset.

Claim: *Liver disease may result in an enlarged liver that can be felt as a lump in the abdomen.*

Evidence Sources:
- **Source 1**: "There are many diseases and conditions that can damage the liver and cause cirrhosis. Some of the causes include Chronic alcoholism..."
- **Source 2**: "Sometimes, liver cysts can become so large that you can feel them through your abdomen. What are the complications of liver cysts?..."
- **Source 3**: "The spleen is a very small organ, usually about the size of a fist. However, many medical conditions, including liver disease and some cancers, can cause the spleen to enlarge..."

Processed Evidence: A variety of other conditions and diseases can lead to cirrhosis, including inflammation and scarring of the bile ducts, called primary sclerosing cholangitis;...Later stages may include jaundice, which is a yellowing of the eyes or skin; bleeding in the gastrointestinal tract; abdominal swelling due to fluid buildup in the abdomen; and confusion or drowsiness....
Machine Label: Supported
Label: Partially Supported

Table 15: Claim for the example from the MEDFACT dataset.

Claim: *Liver disease may cause discomfort or pain in the upper right abdomen.*

Evidence Sources:
- **Source 1**: "Signs of acute liver failure and may include: yellowing of the skin and eyes (jaundice) pain in the upper right abdomen abdominal bulging (ascites) nausea and vomiting general malaise..."
- **Source 2**: "Pain or discomfort in the upper right region of the abdomen. Symptoms that may occur with NASH and cirrhosis (or severe scarring) include..."
- **Source 3**: "Tissue samples show the presence of excess fat in the case of nonalcoholic fatty liver disease; in the case of nonalcoholic steatohepatitis..."

Processed Evidence: Signs of acute liver failure and may include: yellowing of the skin and eyes (jaundice) pain in the upper right abdomen abdominal bulging (ascites) nausea and vomiting generalized feeling of malaise (malaise) disorientation or confusion lethargy breath may have a musty or sweet taste tremor...
Machine Label: Supported
Label: Supported

Table 16: Claim for the example from the MEDFACT dataset.

Claim: *Liver disease may lead to cholestasis, causing itchy skin.*

Evidence Sources:
- **Source 1**: "Diseases of the liver, bile ducts, or pancreas can cause cholestasis. Yellowing of the skin and sclera, itching of the skin, deepening of the color of the urine..."
- **Source 2**: "Pruritus is the most common cutaneous manifestation of liver disease. In patients with liver disease, pruritus is usually associated with cholestasis, such as primary biliary cholangitis, primary sclerosing cholangitis..."
- **Source 3**: "ALGS is characterized by abnormal bile duct development and involvement of extrahepatic organs (e.g., kidneys and eyes), as well as the skeletal and cardiovascular systems. 100% of patients have liver involvement [2, 3], and in addition to jaundice, cutaneous xanthomas, and hepatomegaly, patients present with severe pruritus..."

Processed Evidence: The causes of cholestasis are divided into two categories: intrahepatic causesCauses include acute hepatitis, alcohol-related liver disease, primary biliary cholangitis (with bile duct inflammation and scarring), cirrhosis due to viral hepatitis B or C (also with bile duct inflammation and scarring), certain drugs...
Machine Label: Supported
Label: Supported

Table 17: Claim for the example from the MEDFACT dataset.

Claim: *Liver disease may lead to decreased synthesis of clotting factors and symptoms such as bleeding gums, nosebleeds or bleeding under the skin.*

Evidence Sources:
- **Source 1**: "Bleeding gums or nose is supposed to be a common minor ailment that people mostly don't take seriously, but if you were told that it could be related to liver disease, would you still be able to relax?..."
- **Source 2**: "Nosebleeds are mostly caused by inflammation of the nasal cavity, drying of the nasal mucosa and rupture of capillaries. Nosebleeds in young people may also be related to exertion, exercise and so on. These bleeding is not a big problem, timely treatment can effectively stop bleeding. However, if the bleeding is frequent, large and not easy to stop, it is not so simple, and may indicate other systemic diseases, such as liver disease, blood disease, autoimmune disease, and so on..."
- **Source 3**: "Bleeding in patients with liver disease often manifests itself in a variety of ways; in addition to bleeding from the nose and gums and petechiae on the skin, there may be vomiting of blood or tarry stools..."

Processed Evidence: Why do patients with liver disease have bleeding? This is because a large amount of coagulation factors are synthesized in the liver, and after hepatocellular injury, the function of the liver to produce coagulation factors decreases, followed by a disorder of the coagulation mechanism. In cirrhosis, patients have hypersplenism and increased mechanical destruction of blood, resulting in leukopenia and thrombocytopenia...
Machine Label: Supported
Label: Supported

Table 18: Claim for the example from the MEDFACT dataset.