TCMPHal: A Large-scale Dataset for Hallucination Detection in Traditional Chinese Medicine Pharmacy

Nijia Han[1], **Zimu Wang**[2], **Ziwen Xie**[1], **Wei Wang**[2], **Jia Meng**[1], **John Moraros**[1], **Shuihua Wang**[1,†]

[1]Department of Biosciences and Bioinformatics, School of Science, Xi'an Jiaotong-Liverpool University, Suzhou, China

[2]School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou, China

{Nijia.Han23, Zimu.Wang19}@student.xjtlu.edu.cn,

{Ziwen.Xie, Wei.Wang03, Jia.Meng, John.Moraros, Shuihua.Wang}@xjtlu.edu.cn

Abstract

The rapid proliferation of large language models (LLMs) in medicine highlights their potential to revolutionize research in Traditional Chinese Medicine (TCM). While these models have shown great promise in assisting TCM practitioners by answering herb-related questions, generating syndrome-differentiation reports, and recommending classical formulas, a persistent challenge that arises is the issue of hallucination, where LLMs might produce content that appears plausible yet inaccurate. This issue has received limited attention within the context of TCM research, leaving a significant gap in understanding how hallucination manifests within the unique theoretical frameworks and diagnostic principles. Motivated by this phenomenon, we present TCMPHal, the first dataset specifically curated for hallucination detection in TCM pharmacy, comprising 10,000 high-quality question-answer pairs with hallucination annotations. Our experimental results across diverse LLMs, under standard, knowledge-based, and search engine-augmented conditions, demonstrate the capabilities and limitations of these models. A notable observation is that, for thinking LLMs, incorporating search engine results yields minimal improvement over their intrinsic reasoning abilities. We further conduct an in-depth error analysis, paving the way for future research directions in this domain. We release the TCMPHal dataset at https://github.com/hanninaa/TCMP.

Keywords: Hallucination Detection, Traditional Chinese Medicine Pharmacy, Large Language Models

1. Introduction

Traditional Chinese Medicine (TCM) is progressively expanding beyond its Asian cultural boundaries, gaining widespread global recognition, and serving as a key component of Complementary and Alternative Medicine (CAM; Cohen and Hunter, 2017). However, its unique theoretical framework, diagnostic techniques, and therapeutic principles present considerable knowledge gaps for its cross-cultural popularization. Large Language Models (LLMs), known for their effectiveness in knowledge representation and decision-making (Peng et al., 2023; Ma et al., 2025), as well as their remarkable success in general medical applications (Han et al., 2024; Wang et al., 2025), hold significant potential for advancing the accessibility and integration of TCM (Lu, 2025), such as answering herb-related questions, generating syndrome-differentiation reports, and recommending classical formulas.

Despite their promise, the employment of LLMs, especially in the medical domain, is challenged by the phenomenon of "hallucination," where models may produce seemly plausible yet inaccurate or fabricated information. While recent efforts have been made to detect and mitigate this issue (Pandit et al., 2025; Zuo and Jiang, 2025), this research has focused predominantly on datasets derived from modern medical imaging (Gan et al., 2025), or question-answering datasets (Jin et al., 2019;

Figure 1: Correct and hallucinated examples within the Traditional Chinese Medicine (TCM) scenario, presenting distinct challenges in theoretical constructs and diagnostic logics.

†Corresponding author.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6939–6948
11–16 May 2026.

Chen et al., 2025). These datasets, however, are not suited to capture the distinct theoretical constructs (e.g., *Jun-Chen-Zuo-Shi*) and diagnostic logics in TCM (see Figure 1), creating a significant gap in hallucination detection in this field. Moreover, when constructing datasets for hallucination detection, selecting appropriate references for creating hallucinated examples is critical, as the choice of reference directly determines the semantic plausibility and clinical realism of the synthesized hallucination samples. Inappropriate reference selection, such as pairing conceptually unrelated syndromes or incompatible treatment principles, can lead to trivial or uninformative hallucinations that fail to reflect the nuanced reasoning errors typically made by LLMs in real diagnostic scenarios.

Motivated by this gap, we introduce TCMPHal, the first dataset tailored for **Hal**lucination detection in **TCM P**harmacy. It consists of 10,000 question-answer (QA) pairs, with each question paired with a correct and a hallucinated response. To select appropriate hallucinated references and construct high-quality samples, we build an automated pipeline based on the TCMChat dataset (Dai et al., 2024), as depicted in Figure 2. Given a multiple-choice question alongside both correct and incorrect answers, we construct a *hallucinated* sample by embedding the incorrect answer into the correct answer, thereby introducing subtle inaccuracies using a powerful LLM, GPT-4.1. Conversely, a *correct* sample is derived by expanding the correct answer. This construction method facilitates the evaluation of models and frameworks in detecting hallucinations while ensuring that performance on accurate samples remains uncompromised. We conduct comprehensive experiments on both non-thinking (e.g., GPT-4.1 and DeepSeek-V3 (Liu et al., 2025)) and thinking models (e.g., Qwen3 (Yang et al., 2025) and GLM-Z1) across three different settings: (1) *standard* setting, which directly detects hallucination via prompt engineering, (2) *knowledge-based* setting, which leverages the most relevant domain knowledge, and (3) *search engine-augmented* setting, which retrieves multiple results from search engine tools. Experimental results reveal that existing models fall short in hallucination detection within the TCM pharmacy context, significantly lagging behind human performance, and that search engine results bias thinking models toward making them more prone to classifying non-hallucinated output as hallucinated.

The key contributions of this work can be summarized as follows:

- We present TCMPHal, the first dataset specifically designed for hallucination detection within the TCM pharmacy context, containing 10,000 QA pairs with correct and hallucinated responses.
- We develop an automated pipeline to generate high-quality samples, inserting hallucinations and expanding correct answers based on multiple-choice questions.
- We perform extensive experiments with current LLMs, providing detailed error analysis to underscore their limitations within this domain, offering valuable potential avenues for future model improvements.

2. Related Work

Traditional Chinese Medicine (TCM). With the increased global recognition of TCM, its integration with NLP has seen sustained progress. TCM-SD (Mucheng et al., 2022) centers on syndrome differentiation, a core component of the TCM diagnosis and treatment. TCMBench (Yue et al., 2024) comprises the TCM-ED dataset sourced from the TCM Licensing Exam. TCMChat (Dai et al., 2024) develops an LLM through pre-training and supervised fine-tuning on extensive TCM text corpora and question-answering datasets. While valuable, these datasets and models primarily focus on TCM knowledge, yet they fail to address the hallucination problem, which is widely recognized as one of the most critical challenges in current research.

Medical Hallucination Detection. Early efforts of hallucination detection have primarily concentrated on general-domain texts (Li et al., 2023, 2024). Recently, there has been a shift towards creating specialized hallucination datasets within the medical context. MedHalu (Agarwal et al., 2025) is constructed through various QA pairs, with the focus limited to health queries. MedHallBench (Zuo and Jiang, 2025) leverages medical literature databases to collect scenarios. Med-HALT (Pal et al., 2023) is a multinational dataset from reasoning and memory perspectives, but does not explore its manifestation across diverse medical contexts. Others like MedHal (Mehenni et al., 2025), CMHE (Dou et al., 2024), and MedHallu (Pandit et al., 2025), also evaluate LLMs' ability to detect hallucinations. However, most of these initiatives are centered on modern medicine, with limited attention on TCM. Unlike previous work, we focus on hallucination detection in TCM pharmacy, presenting the first dataset along with a comprehensive analysis of current LLMs' performance.

Hallucination Detection Approaches. Hallucination detection involves the process of recognizing content that is not factual or unverifiable. Recent methods mainly focus on using internal or external knowledge to enhance models' ability. Many efforts leverage large, publicly available knowledge

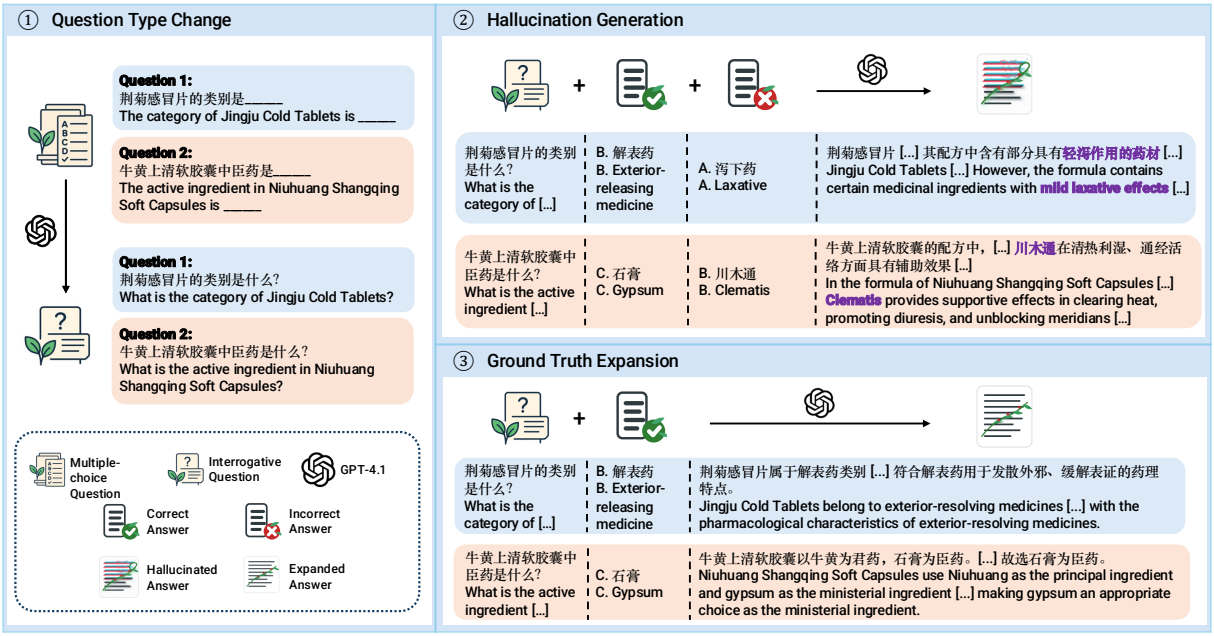

Figure 2: Overview of the dataset construction pipeline, for which we propose deliberate methods for collecting hallucinated (*Hallucination Generation*) and non-hallucinated (*Ground Truth Expansion*) samples.

bases (KBs), such as FAVA (Mishra et al., 2024), or utilize Retrieval-Augmented Generation (RAG; Augenstein et al., 2024; Rakin et al., 2024; Singal et al., 2024) as guidance. However, this approach faces a significant limitation when applied to specialized domains. Specifically, Wikipedia and most other common knowledge bases lack the comprehensive, structured data required for TCM-related knowledge retrieval. To overcome this knowledge deficit and the constraints of static KBs, an increasing number of studies have shifted to deploying search engines (Li et al., 2025a,b). Despite this emerging trend, the highly specialized and sensitive medical domain has been largely underserved. In our work, we build upon these observations and compare the utility of multiple knowledge sources, from relevant knowledge to search engines, so as to bypass the static constraints of conventional KBs and the low efficiency of manual data collection.

3. Dataset Construction

Figure 2 presents the overall pipeline for constructing the TCMPHal dataset, developed through a streamlined three-stage process: *question type change, hallucination generation* and *ground truth expansion*, facilitated by the use of an LLM, GPT-4.1. In this section, we provide a detailed overview of each step.

3.1. Data Generation Pipeline

Data Collection. We begin by collecting TCM QA pairs from the existing TCMChat dataset (Dai et al., 2024), where each question is formulated as a multiple-choice question, complete with an answer and its corresponding explanation. Since incorrect answers are often designed to mislead diagnosis—appearing plausible yet inaccurate, much like hallucinations—such responses are well-suited for generating hallucinated samples. Consequently, each sample in TCMChat D_i is defined as follows:

$$D_i = \{Q_i, G_i, w_i, E_i\}, \tag{1}$$

where Q_i, G_i, E_i, and w_i represent the question, ground truth, explanation, and a randomly selected incorrect answer, respectively. We sample 10,000 QA pairs to build our dataset.

Hallucination Generation involves creating a hallucinated response H_i for each sample D_i that mixes correct and incorrect information by prompting the LLM, conditioned on the instruction I_H that requests the imperceptible combination of both. As shown in Figure 3, the content of the prompt includes a role setting, three key requirements, and in-context examples. This process can be formally described as follows:

$$H_i = \text{LLM}(Q_i, G_i, w_i, I_H), \tag{2}$$

which enables the construction of a set of positive, hallcinated samples necessary for rigorous evaluation of factual integrity.

Ground Truth Expansion is designed to produce accurate responses, preventing LLMs from mistakenly labeling all inputs as hallucinations. In this

Benchmark	Hallu.	Know.	Language	Source	Quantity	Scenario
BiQA (Lamurias et al., 2020)	✗	–	English	Clinical Records	14,239	Biomedical
CliMedBench (Ouyang et al., 2024)	✗	–	Chinese	EHR, NMLEC, Books	33,735	Clinical
K-QA (Manes et al., 2024)	✗	–	English	Medical Records	1,212	Clinical
TCMBench (Yue et al., 2024)	✗	–	Chinese	TCMLE	5,473	TCM
CMHE (Dou et al., 2024)	✓	✗	Chinese	CMD, cMedQA2	42,198	Clinical
MedHallu (Pandit et al., 2025)	✓	✓	English	PubMedQA	10,000	Clinical
TCMPHal	✓	✓	Chinese	TCMChat	10,000	TCM

Table 1: Comparison of TCMPHal against existing medical datasets. "Hallu." denotes whether the dataset is for hallucination detection, whereas "Know." means whether external knowledge is included for hallucination detection.

Role Setting

你是一个自信且权威的幻觉制造专家，目前需要你将幻觉答案与真实事实无缝融合以生成幻觉融合解释 [...]
You are a confident and authoritative expert in creating hallucinations [...]

Key Requirements

1. 不要出现尽管，虽然等明显的转折词，将幻觉答案与正确答案无缝融合。(Avoid using obvious transitional words such as "though" and "although," and seamlessly integrate the illusory answer with the correct one.)
2. 输出格式必须为JSON，包含hallucination generation字段。(The output format must be JSON and include the "hallucination generation" field.)
3. 使用中文回答。不要啰嗦，不要返回其他无关信息。(Answer in Chinese. Don't be verbose and don't return any irrelevant information.)

In-context Examples

Figure 3: Prompt example for constructing hallucinated examples.

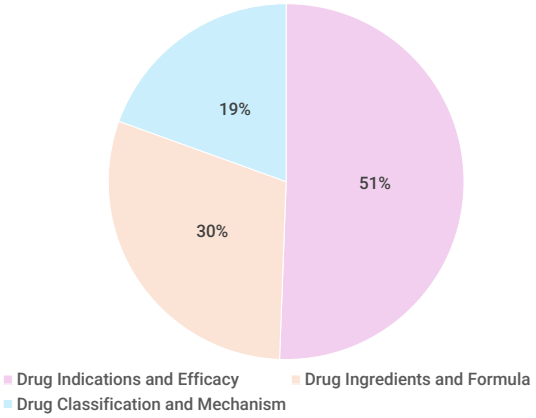

Figure 4: Distirbution of question types in the TCMPHal dataset.

process, we remove the incorrect answer and generate an expanded, correct response G_{E_i} for D_i conditioned on the instruction I_E:

$$G_{E_i} = \text{LLM}(Q_i, G_i, I_E), \quad (3)$$

which establishes a set of negative examples without hallucination. Overall, our methodology utilizes an interrogative question format in TCM principles, combined with carefully constructed positive and negative examples. This strategy effectively mitigates the ambiguity inherent in hallucination detection, relying solely on simple prompting or open-ended generation instructions, thereby ensuring the reliability of the constructed dataset.

3.2. Dataset Characteristics

Overall Statistics. Table 1 presents a comparison of TCMPHal with existing medical datasets across six key dimensions. Notably, TCMPHal distinguishes itself as the first hallucination detection dataset for TCM pharmacy, featuring 10,000 samples with a balanced distribution of correct and hallucinated responses. Besides,the 10,000 questions within our dataset comprehensively cover core knowledge points across the specialized TCM pharmacy domain. The questions are systematically classified into three dimensions using GPT-4.1, as shown in Figure 4: Drug Indications and Efficacy (51%), Drug Ingredients and Formula (30%) and Drug Classification and Mechanism (19%).

Data Quality. To maintain high data quality, we randomly select 200 samples (each with correct and hallucinated instances) for verification by domain experts, with payment and workload agreed upon. Each expert must meet at least two of the following key criteria: a comprehensive understanding of TCM principles, an accredited TCM certification, and extensive practical experience in clinical or practical settings. Prior to involvement, they underwent project-specific training and subsequently assessed the accuracy of each response based on their expertise in TCM principles. The dataset achieved a correctness rate of 97%, indicating the high quality of both the constructed dataset and the underlying data construction pipeline.

Model	Standard	Content	R@5	R@3	R@1	C & R@5	ΔK_C	$\Delta K_{R@5}$
General-Purpose LLMs								
GPT-4.1	0.70	0.83	**0.73**	0.74	**0.73**	0.79	0.13	0.03
GPT-4.1 mini	0.61	0.78	**0.73**	**0.75**	0.70	0.78	0.17	0.12
GLM-4-32B	0.51	0.67	0.67	0.67	0.65	0.71	0.16	**0.16**
DeepSeek-V3	0.64	0.87	0.72	0.72	0.71	**0.80**	**0.23**	0.08
Qwen3-30B-A3B	0.59	0.78	0.71	0.71	0.69	0.75	0.19	0.12
Qwen3-32B	0.64	0.80	0.72	0.72	0.69	0.79	0.16	0.08
Reasoning-Oriented LLMs								
GLM-Z1-32B	**0.73**	0.86	0.72	0.71	0.68	0.76	0.13	-0.01
Qwen3-30B-A3B	0.69	**0.88**	0.72	0.72	0.69	0.79	0.19	0.03
Qwen3-32B	0.70	0.86	**0.73**	0.72	0.70	0.79	0.16	0.03

Table 2: Main evaluation results of LLMs on the TCMPHal dataset. “Standard" refers to detecting hallucinations relying solely on the QA pair and the task instructions. “Content” (C) denotes the incorporation of knowledge content, and “R@i” ($i \in \{1, 3, 5\}$) refers to the inclusion of search engine results. Δ denotes the performance increase under different settings. The best performances are highlighted in bold.

4. Experiments

4.1. Problem Definition

We define the hallucination detection task as follows: Given a question q and a response r generated by an LLM, the objective is to determine whether r contains hallucinated information, which is framed as a binary classification problem. We assess the performance across three different settings: *standard*, *knowledge-based*, and *search engine-augmented* settings.

Standard Setting. In the standard setting, hallucination detection relies exclusively on the QA pair and the task instruction I_S, as described below:

$$R = \text{LLM}(q, r, I_S), \tag{4}$$

where $R \in \{\text{yes}, \text{no}\}$ denotes the hallucination detection result.

Knowledge-based/Search Engine-augmented Settings. To enhance hallucination detection with richer contextual knowledge, we propose integrating two types of supplementary information. First, following MedHallu (Pandit et al., 2025), we incorporate the most relevant knowledge content associated with the question, derived from the explanations in TCMChat. Additionally, we leverage search engine results, incorporating 1, 3, or 5 retrieved results related to the question. Formally, let k denote the knowledge (either from explanations or search results), the task can be defined as follows:

$$R = \text{LLM}(q, r, k, I_S), \tag{5}$$

where $R \in \{\text{yes}, \text{no}\}$ denotes the hallucination detection result.

4.2. Experimental Setup

We conduct comprehensive experiments on both non-thinking and thinking LLMs on the TCMPHal dataset. For non-thinking LLMs, we evaluate GPT-4.1[1], GPT-4.1 mini, GLM-4-32B (Zeng et al., 2024), DeepSeek-V3 (Liu et al., 2025), and Qwen3 models (without thinking ability, Yang et al., 2025). We also experiment with GLM-Z1-32B and Qwen3 models with reasoning abilities to assess the impact of thinking on task performance. The temperature setting varies from 0.2 to 1.3, selected based on the recommended values of different models and the requirements of the current scene. Evaluations are conducted under the accuracy metric, based on the entire dataset, as well as both positive and negative samples.

4.3. Experimental Results

Tables 2 and 3 present the results of LLMs on the TCMPHal dataset, positive examples and negative examples. We also engage two professionals and test human performance on a subset of the positive dataset. From the tables, we have the following observations:

(1) Current LLMs exhibit limitations in detecting hallucinations within TCM pharmacy, with the top-performing model, GLM-Z1-32B, still trailing behind human performance. This can be attributed to the inherent complexity of TCM and the extensive experience accumulated by human experts, underscoring the necessity of the TCMPHal dataset. Among the experimented models, GPT-4.1 and GLM-Z1-32B demonstrate the highest performance across non-thinking and thinking models. Furthermore, while GLM-Z1-32B outperforms GLM-4-32B, the reasoning process does not consistently contribute

[1] https://openai.com/index/gpt-4-1/

Model	Standard	Content	R@5	R@3	R@1	C & R@5
Human Performance on Positive Samples						
TCM Consultant	**0.94**	–	–	–	–	–
TCM Assistant Evaluator	0.85	–	–	–	–	–
General-Purpose LLMs						
	Pos. & Neg.	Pos. & Neg.	Pos. & Neg.	Pos. & Neg.	Pos. & Neg.	Pos. & Neg.
GPT-4.1	0.76 | 0.64	0.84 | 0.82	0.88 | 0.58	0.88 | 0.59	0.88 | **0.57**	0.89 | 0.59
GPT-4.1 mini	0.58 | 0.63	0.80 | 0.76	0.87 | 0.58	0.88 | **0.61**	0.91 | 0.49	0.86 | **0.69**
GLM-4-32B	0.15 | **0.87**	0.46 | **0.87**	0.74 | **0.60**	0.73 | 0.60	0.72 | **0.57**	0.75 | 0.66
DeepSeek-V3	0.52 | 0.75	0.88 | 0.86	0.91 | 0.52	0.91 | 0.52	0.92 | 0.49	0.92 | 0.67
Qwen3-30B-A3B	0.49 | 0.69	0.79 | 0.77	0.88 | 0.54	0.89 | 0.52	0.90 | 0.48	0.86 | 0.63
Qwen3-32B	0.68 | 0.65	0.69 | 0.81	0.93 | 0.48	0.93 | 0.46	0.92 | 0.40	0.91 | 0.56
Reasoning-Oriented LLMs						
GLM-Z1-32B	0.80 | 0.65	**0.91** | 0.81	**0.95** | 0.48	**0.95** | 0.46	**0.96** | 0.40	**0.95** | 0.56
Qwen3-30B-A3B	**0.86** | 0.52	0.90 | 0.86	**0.95** | 0.48	**0.95** | 0.48	**0.96** | 0.41	**0.95** | 0.62
Qwen3-32B	0.81 | 0.59	0.86 | 0.86	0.93 | 0.52	0.94 | 0.50	0.94 | 0.46	0.94 | 0.64

Table 3: Experimental results of LLMs on positive, hallucinated (Pos.) and negative, non-hallucinated (Neg.) samples, respectively. “Standard" refers to detecting hallucinations relying solely on the QA pair and the task instructions. “Content” (C) denotes the incorporation of knowledge content, and “R@i” ($i \in \{1, 3, 5\}$) refers to the inclusion of search engine results.

to improved hallucination detection, as evidenced by the performance comparison of Qwen3 models.

(2) Leveraging external knowledge or search engine results enhances hallucination detection when processing hallucinated samples; however, the search engine results demonstrate interfering effects when it comes to non-hallucinated samples, leading to a noticeable performance degradation by compromising the model’s intrinsic judgment capabilities. By analyzing the performance on non-hallucinated and hallucinated samples, as shown in Table 3, we find that this performance ceiling is largely attributed to the disproportionate impact of non-hallucinated samples, where the degradation in performance on non-hallucinated items offsets the gains achieved elsewhere, particularly among thinking LLMs. We hypothesize that this divergence arises from the adoption of a more conservative decision boundary when the model processes and integrates retrieved information. Specifically, in the Figure 5, the model operating in the standard setting correctly identifies the response as factual and without hallucination. However, when operating under the search engine-augmented setting, the model’s internal thinking process exhibits heightened sensitivity, consequently leading to an increased propensity to misclassify non-hallucination samples as hallucinated.

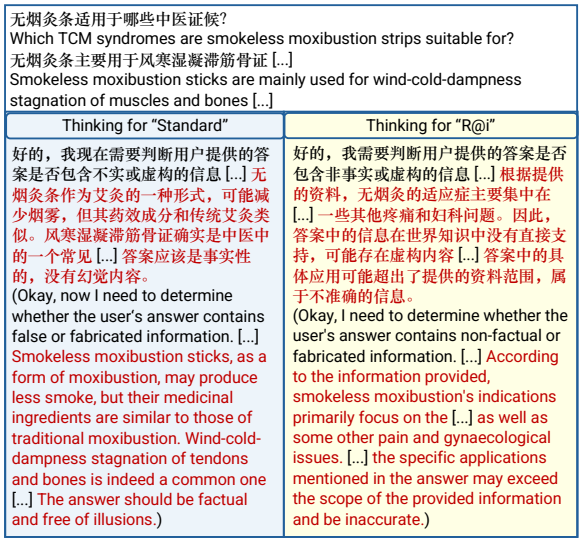

Figure 5: Thinking outputs of Qwen3-32B in standard and search engine-augmented setting, respectively. The model fails to integrate retrieved information effectively, making “hallucinated” conclusion for this non-hallucinated sample.

4.4. Error Analysis

To point out directions for future research, we conduct a thorough error analysis and organize the errors made by current LLMs into the following four aspects, with examples shown in Figure 6:

Efficacy Overgeneralization. Models may be unable to detect overstated or misattributed drug efficacies. For instance, as shown in the upper left part of the Figure 6, the model erroneously classifies a drug as a tonifying medicine. This underscores the lack of deep, critical reasoning capabilities and the absence of real-world, verified clinical trial data, which are essential for distinguishing between valid efficacy claims and exaggerated or unsupported assertions.

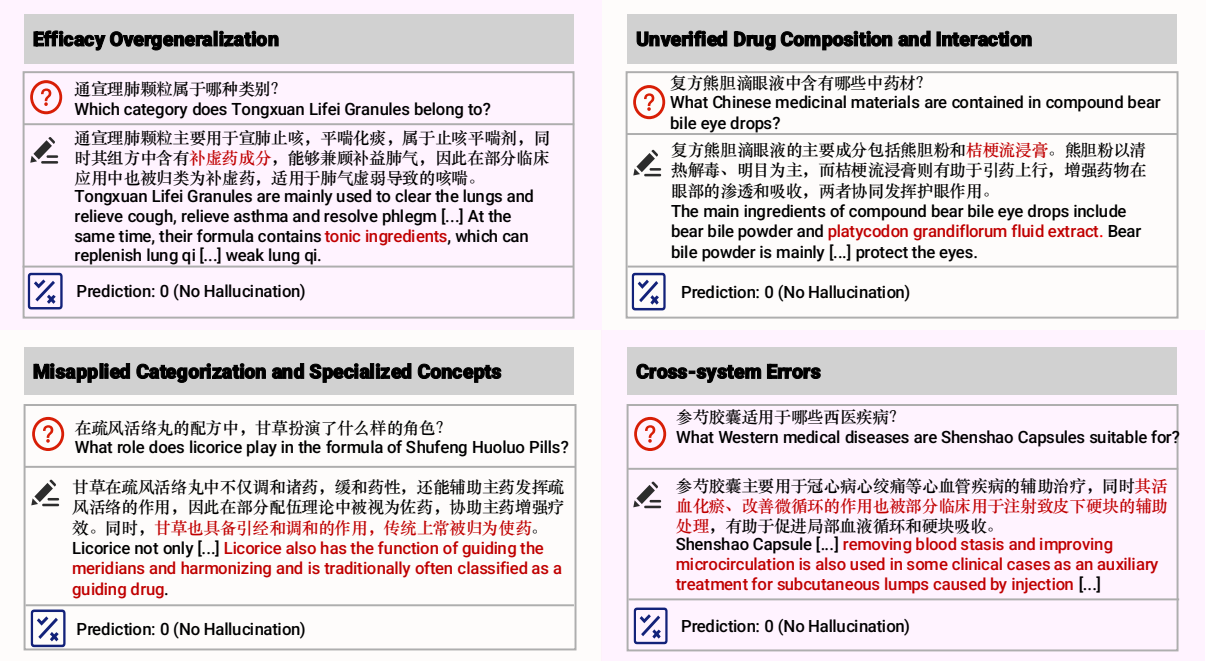

Figure 6: Illustration of the four error types: Efficacy Overgeneralization, Unverified Drug Composition and Interaction, Misapplied Categorization and Specialized Concepts, and Cross-system Errors.

Unverified Drug Composition and Interaction. Current models struggle to identify misapplications of drug composition and interactions. For instance, the model incorrectly includes "*Jie-Geng-Liu-Jin-Gao*" in the upper right part of the Figure 6 as part of the formulation. Identifying erroneous or harmful drug-drug/herb-herb interactions requires more than surface-level recognition but a deep understanding of pharmacological properties and potential conflicts.

Misapplied Categorization and Specialized Concepts. Errors arise when the context surrounding TCM concepts (e.g., *Jun-Chen-Zuo-Shi*), which may involve fabricated or inaccurately attributed medicines. The lower left part of the Figure 6 shows an example that the model incorrectly categorizes a drug, highlighting its inability to align with intricate TCM principles.

Cross-system Errors. LLMs are also difficult to evaluate the use of TCM with the context of Western medicine. As shown in the lower right part of the Figure 6, the response incorrectly recommends TCM herb for a condition based solely on a Western diagnosis, neglecting the essential TCM principle of pattern differentiation. This underscores the inability of LLMs to identify misapplications that stem from the fundamental differences between TCM and Western medical frameworks.

5. Conclusion and Future Work

We introduce TCMPHal, the first dataset specifically designed for hallucination detection in TCM pharmacy, consisting of 10,000 medical QA pairs, with each question paired with a correct and a hallucinated response. Experimental results under standard, knowledge-based, and search engine settings reveal the limitations of current LLMs in detecting hallucinations within the TCM context. Building on this exploration in TCM pharmacy, our future work will focus on developing a more comprehensive dataset with broader aspects of TCM. We will also design specialized models to effectively mitigate hallucinations in the future.

6. Limitations

We recognize the following limitations of our work: (1) Since reasoning-oriented models such as GLM-Z1-32B and Qwen3 models are included, our experiments primarily focused on direct verification methods, without exploring advanced prompting techniques like Chain-of-Thought (CoT). (2) The scope of the dataset is limited to the TCM pharmacy. Future research could expand upon this by constructing more comprehensive datasets encompassing a wider range of topics within TCM.

7. Ethical Considerations

We discuss the following ethical considerations related to our TCMPHal dataset: (1) **Intellectual property.** The original TCMChat dataset is shared under the GPL-3.0 license[2], which is free for research use. (2) **Worker Treatments.** We hired human experts for data verification and human performance with agreed payment and workload, in full accordance with regional legal requirements. (3) **Intended Use.** TCMPHal can be utilized to develop and verify future models in hallucination detection within the TCM pharmacy context. Researchers can also inherit our dataset design to develop their own datasets. (4) **Risk Management.** Since the TCMChat and TCMPHal do not contain private information and the dataset construction process does not require making many judgments about social risks, we believe TCMPHal does not introduce any additional risks. We manually verified some randomly sampled data to ensure the dataset did not contain risky issues.

8. Acknowledgement

We would like to acknowledge the financial support provided by the Postgraduate Research Scholarship (PGRS) (FOSA2406001) at Xi'an Jiaotong-Liverpool University. Additionally, this research has received funding from the Basic Research Program of Jiangsu (BK20241815).

9. Bibliographical References

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, et al. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8):852–863.

Marc Cohen and Jennifer Hunter. 2017. Complementary medicine products: Interpreting the evidence base. *Internal Medicine Journal*, 47(9):992–998.

Yizheng Dai, Xin Shao, Jinlu Zhang, Yulong Chen, Qian Chen, Jie Liao, Fei Chi, Junhua Zhang, and Xiaohui Fan. 2024. Tcmchat: A generative large language model for traditional chinese medicine. *Pharmacological Research*, 210:107530.

[2] https://www.gnu.org/licenses/gpl-3.0.en.html

Hong-Seng Gan, Muhammad Hanif Ramlee, Zimu Wang, and Akinobu Shimizu. 2025. A review on medical image segmentation: Datasets, technical models, challenges and solutions. *WIREs Data Mining and Knowledge Discovery*, 15(1):e1574. E1574 DMKD-00733.R1.

Nijia Han, Zimu Wang, Yuqi Wang, Haiyang Zhang, Daiyun Huang, and Wei Wang. 2024. MTSwitch: A web-based system for translation between molecules and texts. In *Proceedings of the 17th International Natural Language Generation Conference: System Demonstrations*, pages 4–6, Tokyo, Japan. Association for Computational Linguistics.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025a. Search-o1: Agentic search-enhanced large reasoning models.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yongkang Wu, Ji-Rong Wen, Yutao Zhu, and Zhicheng Dou. 2025b. Webthinker: Empowering large reasoning models with deep research capability.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2025. Deepseek-v3 technical report.

Aiping Lu. 2025. Elevating traditional Chinese medicine in global health research: The case for Chinese herbal medicine formulas in mainstream therapeutics. *The Innovation*, 6(4):100811.

Jiayuan Ma, Hongbin Na, Zimu Wang, Yining Hua, Yue Liu, Wei Wang, and Ling Chen. 2025. Detecting conversational mental manipulation with intent-aware prompting. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9176–9183, Abu Dhabi, UAE. Association for Computational Linguistics.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. In *First Conference on Language Modeling*.

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.

Salman Rakin, Md. A. R. Shibly, Zahin M. Hossain, Zeeshan Khan, and Md. Mostofa Akbar. 2024. Leveraging the domain adaptation of retrieval

augmented generation models for question answering and reducing hallucination.

Ronit Singal, Pransh Patwa, Parth Patwa, Aman Chadha, and Amitava Das. 2024. Evidence-backed fact checking using RAG and few-shot in-context learning with LLMs. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 91–98, 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.

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.

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

10. Language Resource References

Vibhor Agarwal, Yiqiao Jin, Mohit Chandra, Munmun De Choudhury, Srijan Kumar, and Nishanth Sastry. 2025. Medhalu: Hallucinations in responses to healthcare queries by large language models.

Tong Chen, Zimu Wang, Yiyi Miao, Haoran Luo, Yuanfei Sun, Wei Wang, Zhengyong Jiang, Procheta Sen, and Jionglong Su. 2025. Medfact: A large-scale chinese dataset for evidence-based medical fact-checking of llm responses.

Chengfeng Dou, Ying Zhang, Yanyuan Chen, Zhi Jin, Wenpin Jiao, Haiyan Zhao, and Yu Huang. 2024. Detection, diagnosis, and explanation: A benchmark for Chinese medial hallucination evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4784–4794, Torino, Italia. ELRA and ICCL.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. 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 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Andre Lamurias, Diana Sousa, and Francisco M. Couto. 2020. Generating biomedical question answering corpora from q&a forums. *IEEE Access*, 8:161042–161051.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10879–10899, Bangkok, Thailand. Association for Computational Linguistics.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Itay Manes, Naama Ronn, David Cohen, Ran Ilan Ber, Zehavi Horowitz-Kugler, and Gabriel Stanovsky. 2024. K-QA: A Real-World Medical Q&A Benchmark. ArXiv:2401.14493 [cs].

Gaya Mehenni, Fabrice Lamarche, Odette Rios-Ibacache, John Kildea, and Amal Zouaq. 2025. Medhal: An evaluation dataset for medical hallucination detection.

Ren Mucheng, Huang Heyan, Zhou Yuxiang, Cao Qianwen, Bu Yuan, and Gao Yang. 2022. TCM-SD: A benchmark for probing syndrome differentiation via natural language processing. In *Proceedings of the 21st Chinese National Conference on Computational Linguistics*, pages 908–920, Nanchang, China. Chinese Information Processing Society of China.

Zetian Ouyang, Yishuai Qiu, Linlin Wang, Gerard de Melo, Ya Zhang, Yanfeng Wang, and Liang He. 2024. CliMedBench: A Large-Scale Chinese Benchmark for Evaluating Medical Large Language Models in Clinical Scenarios. ArXiv:2410.03502 [cs].

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models.

Shrey Pandit, Jiawei Xu, Junyuan Hong, Zhangyang Wang, Tianlong Chen, Kaidi Xu, and Ying Ding. 2025. MedHallu: A Comprehensive Benchmark for Detecting Medical Hallucinations in Large Language Models.

Wenjing Yue, Xiaoling Wang, Wei Zhu, Ming Guan, Huanran Zheng, Pengfei Wang, Changzhi Sun, and Xin Ma. 2024. Tcmbench: A comprehensive benchmark for evaluating large language models in traditional chinese medicine.

Kaiwen Zuo and Yirui Jiang. 2025. Medhallbench: A new benchmark for assessing hallucination in medical large language models. In *Proceedings of The First AAAI Bridge Program on AI for Medicine and Healthcare*, volume 281 of *Proceedings of Machine Learning Research*, pages 205–213. PMLR.