TCM-Align: Curriculum-Aligned MCQ Generation for Traditional Chinese Medicine

Haimo Lu
Tsinghua University
Beijing, China
haimolu@live.com

Li Liu*
Tsinghua University
Beijing, China
liuli95@mail.tsinghua.edu.cn

Jinliang Yuan
Tsinghua University
Beijing, China
yuanjinliang@tsinghua.edu.cn

Yawen Zheng
Tsinghua University
Beijing, China
yvonnetsang16@gmail.com

Zhenyu Wang
Institutes of Science and Development,
Chinese Academy of Sciences
Beijing, China
wangzhenyu@casisd.cn

Kebin Liu
Tsinghua University
Beijing, China
kebinliu2021@tsinghua.edu.cn

Abstract

Traditional Chinese Medicine (TCM) is a comprehensive and historically rich medical system that forms a core part of clinical education in China and beyond. We present a novel framework TCM-Align for automatic generation of multiple-choice QA (MCQs) from Traditional Chinese Medicine (TCM) textbooks. Our approach integrates heuristic-based textbook segmentation, dual-summary semantic filtering, and a MCQ generator to produce high-quality, curriculum-aligned questions. To ensure factual consistency and practical value, we introduce a novel benchmark task specifically designed to evaluate the quality of automatically generated MCQs that encompasses six key evaluation metrics, including a new Source-Consistency Accuracy (SCA) metric. Experiments across seven core TCM subjects yield 2,401 MCQs and demonstrate that our method significantly outperforms a strong baseline (SciQAG) and an ablation variant without the summary filter. Our framework achieves a 94.3% improvement in Explanation Richness and over 10% gains in most other metrics, while maintaining high factual correctness (SCA = 0.98). This work offers a scalable, trustworthy pipeline for generating TCM content—both to support educational applications and to supply the large, high-quality datasets required for model fine-tuning—thereby closing a critical gap in intelligent TCM applications, and lays the foundation for future advances in automated curriculum-level QA generation.

CCS Concepts

• Computing methodologies → Logic programming and answer set programming; **• General and reference → Design**; **• Software and its engineering →** *Software design engineering*.

Keywords

Traditional Chinese Medicine; Large Language Models; Dataset Generation; LLM Hallucination.

[*] Corresponding author

UbiComp Companion '25, Espoo, Finland
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1477-1/2025/10
https://doi.org/10.1145/3714394.3756270

ACM Reference Format:
Haimo Lu, Li Liu, Jinliang Yuan, Yawen Zheng, Zhenyu Wang, and Kebin Liu. 2025. TCM-Align: Curriculum-Aligned MCQ Generation for Traditional Chinese Medicine. In *Companion of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp Companion '25), October 12–16, 2025, Espoo, Finland.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3714394.3756270

1 Introduction

Traditional Chinese Medicine (TCM) is one of the world's oldest medical systems, with a history spanning over 5,000 years and a profound influence on Chinese society and culture [1]. In recent decades, it has gained increasing global recognition, leading to its gradual integration into modern healthcare systems [2]. With the increasing popularity of Large language models (LLMs), such as GPT-4 [3] and the Gemini family [4], in the medical area for more intelligent, scalable solutions in clinical decision support, patient communication, and knowledge management, there is a growing need to explore their potential in TCM applications—such as syndrome differentiation, prescription generation, and intelligent consultation [5].

Despite this growing interest, the performance of general-purpose LLMs in TCM domains remains limited, primarily due to the lack of domain-specific corpora and high-quality annotated datasets [5]. This presents a critical bottleneck, as domain adaptation of LLMs relies heavily on large volumes of relevant and accurate training data [6]. However, the inherent complexity of TCM, including its rich theoretical foundations and specialized terminology, makes the processes of data collection, curation, and validation particularly labor-intensive, typically requiring the involvement of highly trained domain experts [7, 8]. Several Chinese-language TCM datasets have been introduced by collecting existing TCM-related questions with human answers, such as those from the TCM Licensing Exam (TCMLE) and online discussions containing TCM keywords, including TCM-QA [9], TCMBench [7], and TCMD [8]. The volume and breadth of these datasets are limited due to the limitation of selection diversity and fixed QA format. Recent studies have explored the use of LLMs for automated dataset generation, particularly for question–answering tasks, offering a promising approach to expanding existing TCM datasets. For instance, SciQAG [10] proposes LLM-based QA generator and evaluator that takes scientific papers to generate QA pairs filling the demand for more diverse science QA benchmarks. In medical domain, EHR-DS-QA [11] generates a medicial dataset consisting of 150k QA pairs from medical discharge summaries with LLM. These pipelines mostly are not optimized for TCM medical domains. Moreover, many sources of these generated datasets often diverge from formal TCM curriculum system and may suffer from factual inaccuracies.

Therefore, we present a comprehensive framework TCM-Align for automated, curriculum-aligned TCM dataset construction with LLM while maintaining factual consistency. We 1) parse TCM textbooks into chapter-aware segments, 2) apply a dual-summary self-consistency filter to further distill each segment and prune hallucinated or low-confidence content, 3) convert the retained segments into curriculum-aligned multiple-choice questions (MCQs). Our pipeline automatically produces and publicly releases a dataset of over **2,000** high-quality MCQs spanning a wide range of TCM topics, filling a critical gap in TCM corpora resources. We also introduce a novel benchmark task specifically designed to evaluate the quality of automatically generated MCQs that encompasses six key evaluation metrics. Overall, our contributions can be summarized as follows:

- We establish a scalable and quality-controlled framework TCM-Align for TCM dataset generation that lays essential groundwork for introducing LLMs into TCM-related applications.
- We generate a high-quality TCM multiple-choice question (MCQ) dataset that spans diverse TCM topics. This is the first large-scale, curriculum-based QA dataset in TCM, filling a long-standing resource gap and enabling data-driven approaches in TCM.
- We highlight the importance of factual consistency and design a dedicated control mechanism, alongside a benchmark for evaluating the quality of TCM MCQs, addressing the critical need for trustworthy data in TCM applications.

2 Related Work

2.1 TCM Knowledge Datasets

TCM-related question-answer (QA) data is typically only a small subset of large-scale Chinese medical datasets, such as Huatuo-26M [12], which includes 26 million QA pairs derived from online sources. However, these are not TCM-specific. CMExam [13] contains over 60,000 questions sourced from the National Medical Licensing Examination in China (CNMLE), but only 5.8% of them belong to TCM domain.

In response to this limitation, several specialized TCM QA datasets have been compiled. Table 1 provides a comparative overview of these resources. For instance, TCM-QA consists of 801 QA pairs extracted from BaiduWenKu and subsequently verified by domain staff [9]. TCMBench leverages QA items from the TCM Licensing Exam (TCMLE), with human experts involved in the selection and screening process [7], though it is not yet publicly available.

2.2 LLM-based QA Generation

Broadly, LLM-based QA generation can be categorized into two domains: 1) Hybrid human-LLM pipelines. Kotschenreuther et al. [11] generate questions from medical discharge summaries in MIMIC-IV-Note using LLM but still require human experts to inspect the outputs. 2) Fully-Automated LLM-based pipelines. SciQAG [10] feeds each entire scientific paper to an LLM, produces a fixed set of ten questions per paper, and is thus potentially constrained by article length. MCQGen [14] combines LLM with retrieval-augmented generation (RAG) and advanced prompt engineering like chain-of-thought and self-refine prompting for generation. However, it critically depends on a predefined external knowledge base, limiting its general applicability to new domains without prior data preparation.

By contrast, TCM-Align pre-segments TCM textbooks along their native hierarchy and carries out the entire segment summarization, dual-summary filtering, and MCQ generation pipeline fully automatically as shown in Figure 1, eliminating the need for manual intervention while achieving finer-grained topical coverage.

2.3 Evaluation Metrics

Automatic assessment of generated QA material spans three tiers. **Traditional lexical-overlap scores**, such as ROUGE [15] and BLEU [16], measure surface similarity of two texts but say little about factual correctness or reasoning depth. **Semantic metrics**, such as BERTScore [17], understands contextual meaning of given texts and improves correlation with human judgement, but simple semantic metrics cannot capture the different aims between source texts and generated contents. **Treat a strong LLM itself as an evaluator**. GPTScore [18] prompts the model to assign holistic or rubric-based ratings and report markedly higher agreement with experts than overlap metrics. Domain-specific rubrics like RACAR in SciQAG [10] grade and filter their generated QA pairs using a single five-factor score. We build on this LLM-as-judge paradigm to introduce six fine-grained dimensions tailored to multiple-choice TCM questions.

3 Proposed Framework

3.1 Pipeline Overview

TCM-Align automates question generation from TCM textbooks via several key steps: (1) Heuristic-based Knowledge Extraction, (2) Dual-Summary Generation with semantic filtering, (3) Multiple Choice Question Generation with a format verifier.

3.2 Heuristic-based Knowledge Extraction

We designed a heuristic-based knowledge extraction method to extract knowledge from the TCM textbooks. The key idea is that a textbook is organized and designed in a way that knowledge are clustered into different segments of texts that are semantically similar, and that could serve as a good unit of knowledge. We use the chapter, section, and sub-section structure of the textbook as our heuristics to extract knowledge from the textbook. Users of this framework should modify the heuristics extraction rules to fit their own needs based on the structures of their inputs.

3.3 Dual-Summary Generation

While the extracted segments generated from the previous step provide a structurally meaningful unit of knowledge, they often contain multiple strands of information, making it challenging for question generation models to directly identify the core concepts. To address this, we propose a summary generation step that distills each segment into a focused semantic representation, enabling more effective and trustworthy question generation. As inspired by Nenkova et

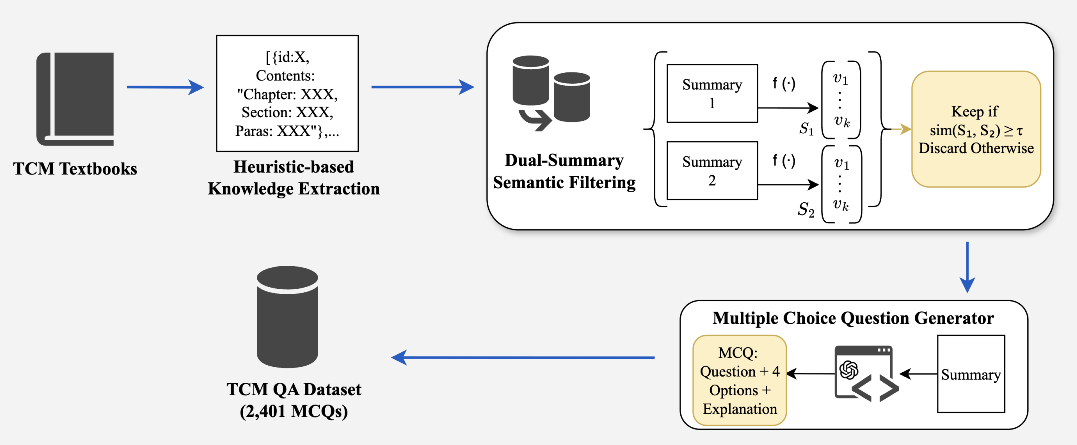

Figure 1: Overview of our proposed framework for TCM question-answer generation.

Dataset	Size	Open-source?	Automated Dataset Creation?
TCM-QA [9]	574	Yes	No
TCMBench [7]	5473	Claims Yes but not yet	No
TCMD [8]	3451	No	No

Table 1: Summary of recent Chinese TCM QA datasets

al. [19], when a content appears across multiple summaries from human experts, it is more likely to be a core concept. Building on this, we prompt the LLM to produce two independent summaries for every segment. If the two summaries converge on the same information—yielding high semantic similarity—we infer that this shared content represents the segment's central idea; if they diverge, it signals that the segment's salient points are unclear or that hallucinations may be present [20], hence we discard that segment. While two samples give a coarse signal, we find the gain in speed outweighs the loss in reliability.

Since comparisons are required for each summary pair, we need to eliminate external factors that could influence similarity outcome, for example, formats and structures of generated summaries could be very different. Thus, we design the prompt to instruct our used model to output summaries in json format, with fixed structures and entry names specially tailored for TCM topics.

Given a textbook segment T_i, we prompt an LLM to generate two summaries: $S_i^{(1)}$ and $S_i^{(2)}$. To assess their semantic consistency, we encode both summaries into vector representations using an embedding function $f(\cdot)$ inspired by [21]:

$$\mathbf{e}_i^{(1)} = f(S_i^{(1)}), \quad \mathbf{e}_i^{(2)} = f(S_i^{(2)}) \tag{1}$$

We then compute the similarity between the two embeddings:

$$\text{sim}(S_i^{(1)}, S_i^{(2)}) = \frac{\mathbf{e}_i^{(1)} \cdot \mathbf{e}_i^{(2)}}{\|\mathbf{e}_i^{(1)}\| \cdot \|\mathbf{e}_i^{(2)}\|} \tag{2}$$

If $\text{sim}(S_i^{(1)}, S_i^{(2)}) \geq \tau$, where τ is a predefined threshold (where we set $\tau = 0.92$), we consider the summary pair semantically aligned. One summary is retained as $\hat{S}_i$ for subsequent MCQ generation. Otherwise, T_i is discarded from further processing:

$$\hat{S}_i = \begin{cases} S_i^{(1)} & \text{if } \text{sim}(S_i^{(1)}, S_i^{(2)}) \geq \tau \\ \text{discard} & \text{otherwise} \end{cases} \tag{3}$$

This approach reduces the likelihood of hallucinated or inconsistent content being passed downstream, and serves as a filtering mechanism to enforce semantic agreement between LLM-generated outputs to ensure the quality of the later generated MCQs.

3.4 Multiple Choice Question Generator

Taking the retained summaries as source, We designed a prompt using LLM to generate multiple choice questions. Each question is generated with one main question text, and with four options, where only one option is correct. Each MCQ also come with an explanation, which complements the question to provides a better understanding. We then designed a MCQ format verifier to perform shallow verification that each generated MCQ satisfies the following structural criteria:

- There is a question text, and four options.
- Only one option is correct.
- There is an explanation accompanying the question.

4 Dataset

4.0.1 Textbook Collection. We collected seven TCM textbooks which belong to Chinese National TCM Higher-Education Planned Textbooks series as our generation resources. These textbooks form a cohesive, Chinese ministry-endorsed series that covers the core disciplines of TCM. For all resources we omitted any images and tables through extracting the plain text, forming raw data that covers an extensive range of TCM topics, from basic theories to clinical practices, offering a comprehensive, domain-trusted corpus on which to build and evaluate our framework.

4.0.2 Dataset Generation. To accommodate the layout conventions of our textbook corpus, we first analyzed each volume's structure and derived a set of regular-expression heuristics that recognize **Chapter**, **Section**, and **Subsection** headings. These heuristics drive the knowledge-extraction pipeline described in Section 3.2, ensuring that every extracted segment respects the textbooks' native hierarchy and semantic boundaries.

We then adopt the framework to generate Summaries and then MCQs for each textbook to get a dataset of 2401 MCQs, when generating summary, we set a threshold of 0.92 for the cosine similarity of the two summaries, any summary pair with similarity less than 0.92 were discarded.

4.1 Dataset Statistics

The number of MCQs generated for each TCM category is shown in Figure 2.

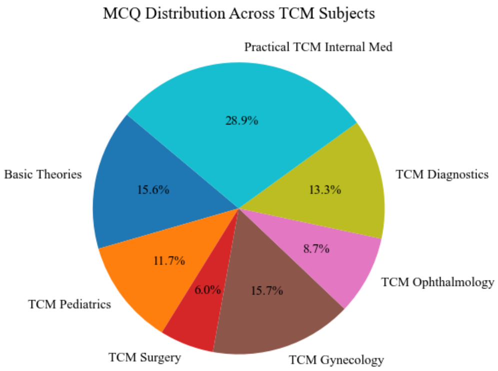

Figure 2: Distribution of MCQs generated per TCM category. Total MCQs: 2,401.

5 Experiments

5.1 Experiment Setup

Baseline vs. TCM-Align We adopt SciQAG [10] as a strong baseline. SciQAG framework is an fully automated QA generation framework that implements a filtering module to generate high-quality QA datasets from a given source text. We sampled 50 MCQs per textbook for the two methods, yielding 350 MCQs each. For the SciQAG baseline we keep the authors' default hyper parameters with a single change: their question-generation prompt is revised to output multiple-choice questions instead of open QA pairs. Consistent with their original protocol, we feed ten consecutive textbook segments as one article input to SciQAG to make sure we generate same number of questions with same number of segment input, yielding a fairer comparison.

Ablation: Summary vs. No Summary To isolate the contribution of dual-summary self-consistency filter, we run our pipeline without the summary generation stage. Specifically, we feed the raw textbook segment obtained from knowledge extraction directly into the MCQ generator and skip cosine-similarity filtering. We then sampled 50 MCQs per textbook, producing another 350 items as ablation set for evaluation. Each evaluation set is then passed through metrics for evaluation. For SCA, regarding time and cost concerns, we sampled 30 MCQs per textbook, producing another 210 items for evaluation.

Implementation Details Unless stated otherwise, all prompts are executed with GPT-4.1 via the OpenAI API. We deliberately choose this larger model rather than the GPT-4.1-mini variant employed inside our pipeline to ensure that evaluation-time decisions (e.g. SCA scoring) are made by a model that is at least as capable as the generators, thus avoiding an evaluator handicap. We ran each question through our defined metrics then average the scores of each question to get the final score.

5.2 Metrics

Source-Consistency Accuracy (SCA) For each MCQ $\langle q, \{o_i\}, o^* \rangle$ we prompt an independent LLM instance with the original paragraph S plus q and its four options without revealing o^*:

Based on the source text, identify the correct answer for the following question.

- Source text: [S]
- Question: [q]
- Options: [o_1, o_2, o_3, o_4]

Let $\hat{o}^*$ be the option selected by the model.

$$\mathrm{SCA} = \frac{1}{N}\sum_{j=1}^{N} \mathbb{I}\left[\hat{o}_j^* = o_j^*\right] \in [0,1]. \tag{4}$$

where N is the number of MCQs in the evaluation set and $\mathbb{I}[\cdot]$ is the indicator function.

Intuitively, SCA measures, for each MCQ j, whether an external LLM—given the same evidence (namely the original source paragraph S)—converges on our framework's provided answer o_j^* (generated from the summary). We then average these $0/1$ agreement indicators over all N questions. A higher SCA indicates that a larger fraction of questions have the correct option indeed entailed by the source text, echoing the self-consistency rationale of Wang et al [22]. Concretely, the two reasoning paths we compare are:

- The summary-level chain of thought $\tilde{S}$ used by TCM-Align to derive $\langle q, \{o_i\}, o^* \rangle$
- The full-text chain of thought S that an independent LLM follows when selecting $\hat{o}^*$.

If these independent paths converge on the same answer ($\hat{o}^* = o^*$), we count the item as consistent while divergence signals a potential factual error in the generated MCQ.

Reasoning Span Score (RSS) assesses whether answering requires local (single-sentence) information from the source text or integration across adjacent contents. A higher score indicates a more global question. We would want our generated MCQs utilize as much information of the source text as possible, which means higher RSS.

Explanation Consistency (EC) quantifies how well the explanation supports the marked correct option o^*. A higher EC indicates that the explanation explicitly endorses o^*, gives evidence from the source text, and contains no statements that contradict o^* or endorse a distractor option.

Explanation Length Quality (ELQ) assesses the appropriateness of the explanation length and information density. Higher scores indicate concise explanation with less redundancy.

Distractor Quality (DQ) measures how relevant the distractor options are to the question itself, and whether ruling them out would require reasoning instead of just using superficial cues. Higher scores indicate distractors that are contextually appropriate and require genuine understanding to distinguish from the correct answer. A good MCQ should have plausible yet ultimately incorrect distractor.

Explanation Richness (ER) A strong explanation justifies the whole option set, and brings more educational purposes. This metric evaluates whether the explanation states why the correct answer is right and why distractors are wrong. A higher score indicates all options are well-explained.

We designed these metrics around the idea of what a MCQ with rich information signal would look like, to serve as an evaluation standard for both our generated MCQs, and when we compare with other works. Scores for each metric ranges from 0 to 10, and the higher the score, the better the MCQ. Criteria for each score range are provided in details for better precision.

6 Evaluation

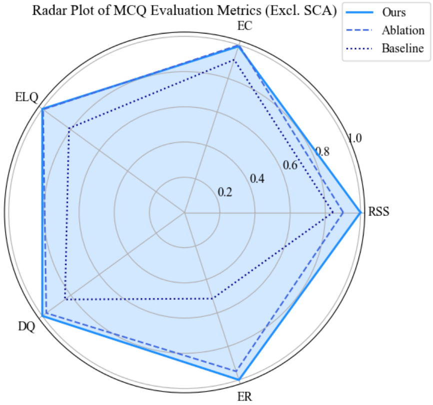

Figure 3: Performance comparison of TCM-Align against Ablation, and Baseline approaches on MCQ evaluation metrics, SCA excluded due to different scoring range.

6.1 Overall Comparison

The results of the six-metric comparison with the baseline are shown in Table 2. Across all metrics, TCM-Align outperforms the baseline, with the largest absolute gain in ER at 94.3%, and over 10% improvement in most other dimensions. This demonstrates that TCM-Align generates MCQs whose answers require more comprehensive reasoning over the source content(RSS). The explanations produced are more aligned with the correct option(EC), exhibit higher information density with less redundancy(ELQ), and are more likely to express not only why the correct answer is correct, but also why the distractors are incorrect(ER). In addition, the distractors themselves are more contextually appropriate and demand a deeper understanding to differentiate from the correct choice(DQ).

6.2 Ablation Study

Dual-summary self-consistency filter. To isolate the contribution of the filter, we re-ran the pipeline without it. Table 2 and Table 3 show that the full framework improves five of the six metrics, with gains of +10.9% in RSS and +5.4% in ER. The only metric that favours the ablation is EC, by a marginal 0.8%. The marked increase in RSS confirms our hypothesis that distilling each segment into a filter-certified summary encourages the LLM to reason over a broader portion of the source text rather than anchoring on local snippets. The simultaneous gains in ER, DQ, and ELQ indicate that the filter denoises the input, allowing the generator to capture more contextually relevant details while preserving factual correctness with SCA unchanged at 0.98.

Table 2: Quantitative Comparison of MCQ Generation Methods

Metric	TCM-Align	Ablation	Baseline
SCA	**0.98**	0.97	**0.98**
RSS	**4.48**	4.04	3.78
EC	8.58	**8.65**	7.89
ELQ	**7.68**	7.64	6.24
DQ	**6.10**	5.93	5.13
ER	**6.47**	6.14	3.33

Table 3: Percentage improvement of TCM-Align over Ablation and Baseline (higher is better)

Metric	vs. Baseline (%)	vs. Ablation (%)
SCA	0.0	1.0
RSS	18.5	10.9
EC	8.8	–0.8
ELQ	23.1	0.5
DQ	19.0	2.9
ER	94.3	5.4

Heuristic-based knowledge extraction. When the same ChatGPT model is provided with the full source paragraph and answers each MCQ, it agrees with the labeled answer 97–98% of the time (Table 2). The near-ceiling scores confirm that even the baseline's factual grounding is strong with heuristic-based knowledge extraction and our qualitative improvements are not achieved by sacrificing correctness.

The ablation study confirmed the effectiveness of TCM-Align, of which the heuristic-based knowledge extraction can deliver accurate result, and the summary generation with filtering further improves the generation quality.

7 Conclusion

In this work, we proposed a fully automated semantic QA generation framework tailored to TCM textbooks. Our method integrates chapter-aware heuristic extraction, dual-summary semantic filtering, and high-precision MCQ generation to build a large-scale, curriculum-aligned dataset.

Through comparative experiments against a strong baseline and an ablation variant without the summary filtering step, we demonstrated that our framework consistently outperforms alternatives across five key quality metrics while maintaining strong factual correctness, as confirmed by a high SCA of 0.98. Notably, ER improved by 94.3% compared to the baseline, highlighting the educational depth introduced by our pipeline. The dual-summary filter was shown to play a crucial role in enhancing contextual grounding and de-noising the input for better question generation. Our framework yields a high-quality dataset of 2,401 TCM MCQs spanning seven core domains, providing an open resource to support downstream applications in education, evaluation, and curriculum development.

8 Acknowledgement

The authors would like to thank the anonymous reviewers and Guang'anmen Hospital, China Academy of Chinese Medical Sciences. This work was supported by the National Natural Science Foundation of China (Grant No. 62441233).

References

[1] Luís Carlos Matos, Jorge Pereira Machado, Fernando Jorge Monteiro, and Henry Johannes Greten. Understanding traditional chinese medicine therapeutics: An overview of the basics and clinical applications. *Healthcare (Basel)*, 9(3):257, March 2021.

[2] Aisen Xu. Traditional chinese medicine in modern healthcare: A bridge between ancient wisdom and contemporary practice. *International Journal of Education and Humanities*, 12(1):69–72, January 2024.

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[4] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[5] Heyi Zhang, Xin Wang, Zhaopeng Meng, Zhe Chen, Pengwei Zhuang, Yongzhe Jia, Dawei Xu, and Wenbin Guo. Qibo: A large language model for traditional chinese medicine. 2024.

[6] Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. 2023.

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

[8] Ping Yu, Kaitao Song, Fengchen He, Ming Chen, and Jianfeng Lu. Tcmd: A traditional chinese medicine qa dataset for evaluating large language models. *arXiv preprint arXiv:2406.04941*, 2024.

[9] Li Yizhen, Huang Shaohan, Qi Jiaxing, Quan Lei, Han Dongran, and Luan Zhongzhi. Exploring the comprehension of ChatGPT in traditional chinese medicine knowledge. 2024.

[10] Yuwei Wan, Yixuan Liu, Aswathy Ajith, Clara Grazian, Bram Hoex, Wenjie Zhang, Chunyu Kit, Tong Xie, and Ian Foster. SciQAG: A framework for auto-generated science question answering dataset with fine-grained evaluation. 2024.

[11] Konstantin Kotschenreuther. EHR-DS-QA: A synthetic QA dataset derived from medical discharge summaries for enhanced medical information retrieval systems, 2024.

[12] Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. Huatuo-26M, a large-scale chinese medical QA dataset. 2023.

[13] Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, and Michael Lingzhi Li. Benchmarking large language models on CMExam – a comprehensive chinese medical exam dataset. 2023.

[14] Ching Nam Hang, Chee Wei Tan, and Pei-Duo Yu. MCQGen: A large language model-driven MCQ generator for personalized learning. *IEEE Access*, 12:102261–102273, 2024.

[15] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[16] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, Morristown, NJ, USA, 2001. Association for Computational Linguistics.

[17] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. 2019.

[18] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. GPTScore: Evaluate as you desire. 2023.

[19] Ani Nenkova and Rebecca Passonneau. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA, 2004. Association for Computational Linguistics.

[20] Potsawee Manakul, Adian Liusie, and Mark J F Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. 2023.

[21] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[22] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Nathan Scales, and *et al.*. Self-consistency improves chain of thought reasoning in language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.