Supplementary Information for: [TCMKG: Knowledge-Grounded Question Answering in Traditional Chinese Medicine via Dynamic Prompt Evolution Framework]

1 Inter-Annotator Agreement Calculation

To evaluate the consistency between the two experts before adjudication, this study used **Cohen's** κ as the inter-annotator agreement metric. This measure accounts for chance agreement beyond the observed agreement and is defined as:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{1}$$

Where P_o denotes the observed agreement and P_e denotes the expected agreement by chance.

For each entity type t, all entities annotated by the two experts across the full set of medical records were first collected, and their union was used to construct a candidate entity set V_t. Then, for each medical record i and each candidate entity $e \in V_t$, a binary decision was assigned for each annotator: 1 if the annotator labeled entity e as belonging to type t in record i, and 0 otherwise. In this way, each "record–candidate entity" pair was converted into one binary annotation decision.

Let n_{11}, n_{00}, n_{10}, and n_{01} denote the numbers of cases where both annotators assigned 1, both assigned 0, annotator A assigned 1 while annotator B assigned 0, and annotator A assigned 0 while annotator B assigned 1, respectively. The total number of binary decisions is

$$n = n_{11} + n_{00} + n_{10} + n_{01} \tag{2}$$

The observed agreement is then calculated as:

$$P_o = \frac{n_{11} + n_{00}}{n} \tag{3}$$

The expected agreement by chance is calculated as:

$$P_e = P(A = 1)P(B = 1) + P(A = 0)P(B = 0) \tag{4}$$

$$P(A = 1) = \frac{n_{11} + n_{10}}{n}, \quad P(B = 1) = \frac{n_{11} + n_{01}}{n} \tag{5}$$

$$P(A=0)=\frac{n_{01}+n_{00}}{n},\quad P(B=0)=\frac{n_{10}+n_{00}}{n} \tag{6}$$

Substituting P_o and P_e into the formula above yields Cohen's κ for that entity type. The overall Cohen's κ was computed by merging all binary decisions across the seven entity types, so as to reflect the overall agreement between the two experts.The agreement between the two experts' independent annotations prior to arbitration on 200 medical records is shown in Table 1.

2 Entity Annotation Guidelines

See Table 2

3 Prompt Design Framework

During the prompt iteration process, we found that it is crucial to explicitly specify the instructions and tasks expected of the model. The more descriptive and detailed the prompt, the better the model's performance. In our prompt design, we use clear delimiters `'###'` to separate instructions from the input context, which helps the large language model more effectively execute the task. An example of the prompt framework for prescription (formula) extraction is shown in the Fig. 1.

4 Proof of the Performance Gains from LLM Error Analysis

To examine whether 'LLM Error Analysis' in the construction of Prompt3 yields attributable incremental benefits for the entity extraction task, we introduce an additional ablation study to verify the contribution of 'LLM Diagnosis' to prompt optimization.

The procedure is as follows. We start from the outputs produced by the Prompt2 template on the entity-type classification and extraction task. Domain experts then select representative instances from the error samples of each entity type, and ultimately construct an ablation evaluation set containing 43 samples. Among them, there are 7 samples of the `Etiology & Pathogenesis` type, and 6 samples for each of the other six entity types. All experiments are conducted on the exact same evaluation set and evaluation script, comparing the following three conditions:

1. **Prompt2**: directly perform the task using the Prompt2 template;
2. **Prompt2 + LLM-assisted error summarization**: based on Prompt2, introduce an LLM to categorize and attribute errors on the evaluation samples and to diagnose their causes, and then adjust the prompt accordingly;
3. **Prompt2 + expert error summarization**: based on Prompt2, experts consolidate and summarize the LLM diagnostic outputs and revise the prompt.

We use F1 as the primary metric because Precision and Recall in entity extraction often involve a trade-off, and reporting only P or R may introduce bias. F1 jointly reflects false positives and false negatives, making it suitable for a comparable

incremental assessment of the overall gains brought by different prompt optimization strategies in an ablation setting. Meanwhile, we still report the complete P/R metrics for each condition in the tables below. Tables 3, 4, and 5 present Precision/Recall/F1 for the three conditions across the seven entity types; Table 6 summarizes the improvement brought by adding error analysis relative to Prompt2, as well as the net increment of expert error analysis relative to LLM error analysis, in order to validate the incremental value of each form of error analysis in prompt optimization. The overall ablation results reported here serve only as evidence of the effectiveness of the error analysis module. The magnitude of the overall improvement is not consistent with the results obtained on the 200-sample test set used in the main experiment.

5 Entity Extraction Results of Baichuan4-Turbo

See Table 7.

6 Entity Extraction Results of DeepSeek-V3

See Table 8

7 Entity Extraction Results of GLM-4-Plus

See Table 9

8 Entity Extraction Results of Qwen-Max

See Table 10

9 TCMKG QA Error Analysis

Our TCMKG question-answering framework consists of three sequential stages: (1) generating a Cypher query statement from the input question; (2) retrieving supporting evidence from the Neo4j graph database by executing the query; and (3) generating the final answer based on the retrieved results. Through concrete empirical validation in conjunction with the overall question-answering framework, we found that QA errors can be categorized into two types: retrieval failures and generation failures, as illustrated in Table 11.

Retrieval failure refers to cases in which the system fails to obtain the correct supporting evidence from the knowledge graph. In this framework, retrieval is not achieved through conventional document recall; rather, it is accomplished through large-language-model-based Cypher generation and graph query execution. As shown in Question 1, the context entities retrieved by TCMKG include not only symptoms but also several disease entities. Based on the question and the ground-truth answer, we can infer that, during question-driven retrieval, the model inserted inappropriate constraints into the query template, thereby leading to erroneous returned results. In addition, the current template mainly retrieves one-hop neighborhood information

centered on candidate nodes. This design is effective for simple factual questions, but it is often insufficient for questions that require multi-hop reasoning or evidence aggregation across multiple related entities. For example, for questions such as "Which prescription corresponds to a given disease under a certain therapeutic method and treatment principle?", the model is unable to fully obtain the evidential chain needed to support the answer.

Generation failure refers to cases in which the system has already retrieved relevant evidence but still fails to generate the correct final answer. As shown in Question 2, under the current question, the contextual retrieval results have already returned the complete target entities. However, the model omitted one entity when producing the answer, resulting in an incomplete or inaccurate response.

Overall, this analysis indicates that future improvements need to address both stages of the QA pipeline simultaneously. At the retrieval stage, more robust entity alignment, schema-constrained Cypher generation, and multi-hop query design are expected to improve the completeness and accuracy of evidence acquisition. At the generation stage, more explicitly structured retrieval results, together with stronger constraints on evidence faithfulness or abstention mechanisms, may reduce the production of unsupported or incomplete answers.

10 Data and code availability

The raw data, source code, and experimental configurations supporting the findings of this study are publicly available in the following GitHub repository: https://github.com/NobodyDoinb/TCMKG.

Table 1 Inter-annotator agreement measured by Cohen's κ for the two experts' independent annotations before adjudication. The table reports the agreement for each entity type as well as the overall agreement across all entity types.

Entity Type	**Cohen's κ**
Symptoms	0.7025
Pulse Diagnosis	0.2658
Tongue Diagnosis	0.6893
TCM Herbs	0.9253
TCM Formulas	0.8723
Etiology & Pathogenesis	0.4627
Treatment Principles	0.2999
Overall	0.7059

Note: Cohen's κ was computed from the two experts' independent annotations before adjudication. Higher values indicate stronger agreement between annotators after accounting for chance agreement.

Table 2 Entity Types and Annotation Rules

Entity Type	Entity Concept	Entity Annotation Standards
Symptoms	Abnormal sensations and states manifested by the body due to disease occurrence, such as "cough", "abdominal pain", "fever"	1. Follow the principle of minimal semantic units. When it does not affect the meaning of the text, the annotated subject should be the smallest possible word unit. 2. Directional words and time words should be marked. Degree adverbs and modifiers are not annotated. For example, "right hand and foot numbness", "lighter in the afternoon", "fever in the afternoon" should be annotated as a whole. "Night fever with chills" should be marked as "night fever", "persistent fever" should be marked as "persistent fever". 3. For multiple symptoms with parallel relationships, follow the independent annotation rule. For progressive relationships, use the holistic annotation principle. For example, "dizziness and vertigo" should be marked as "dizziness" and "vertigo", while "heat with fullness" should be annotated as a whole.
Tongue Diagnosis	Tongue body morphology, color, tongue quality and coating color	1. Follow the principle of minimal semantic units. When it does not affect the meaning of the text, the annotated subject should be the smallest possible word unit.
Pulse Diagnosis	Pulse conditions perceived through pulse-taking (radial pulse diagnosis)	1. Follow the principle of minimal semantic units. When it does not affect the meaning of the text, the annotated subject should be the smallest possible word unit. 2. Pulse positions and modifiers are not annotated. For example, "weak pulse, more pronounced at the right guan" should be marked as "weak pulse".
Etiology & Pathogenesis	Causes of disease occurrence and the mechanisms of disease development, progression, and changes	1. Follow the principle of minimal semantic units. When it does not affect the meaning of the text, the annotated subject should be the smallest possible word unit. 2. Annotations should include changes in etiology, pathological nature, disease location, disease tendency and their mechanisms, such as "stomach qi moving in the field", "stomach qi not drumming", "qi rushing upward irregularly", "wind attacking the yangming".
Treatment Principles	Basic principles that must be followed in treating diseases and specific methods and measures for disease treatment	1. Follow the principle of minimal semantic units. When it does not affect the meaning of the text, the annotated subject should be the smallest possible word unit, such as "harmonizing and warming with heat-clearing and detoxifying" should be marked as "harmonizing and warming" and "heat-clearing and detoxifying".
TCM Formulas	Chinese medicine prescriptions composed of multiple herbs according to compatibility principles and clinical experience. Formula names follow the format of "formula name + dosage form", including "XX decoction", "XX pill", "XX formula", etc.	1. Follow the principle of minimal semantic units. When it does not affect the meaning of the text, the annotated subject should be the smallest possible word unit. 2. Annotate the specific prescription names mentioned in medical cases as standard, such as "Erchen decoction", "Sangju pill". 3. For formulas that specify both formula name and dosage form, annotate as a whole including dosage and quantity, such as "five doses of black medicine warming formula", "three doses of black medicine stomach pain formula".
TCM Herbs	Herbs used in prescriptions, guided by basic TCM theory and Chinese medicine processing methods, summarized through long-term practice for disease prevention and treatment, and health maintenance	1. Follow the principle of minimal semantic units. When it does not affect the meaning of the text, the annotated subject should be the smallest possible word unit. 2. Processing methods of Chinese medicine should be annotated together with the herb, such as "stir-fried white peony", "ophiopogon wine drink", "gleditsia stir-fried yellow". 3. Decoction methods for Chinese medicine are not annotated. For example, in "astragalus decocted first", only "astragalus" is marked.

###Role
你是一个中医的方剂抽取大模型。
You are a large language model specialized in extracting Traditional Chinese Medicine formulas.

###Task
你的任务是提取出输入文本中最终用于治疗的方剂，用列表的格式给出，若文本中无方剂请输出null。
Your task is to extract the final treatment formulas (prescriptions) mentioned in the input text. Return them in a list format. If no formulas are present, output null.

###Definition
方剂：根据配伍原则和医师的临床经验，以若干药物配合组成的中医处方，方剂的名称多为“方名+剂型”的形式，包括“XX 汤”“XX 丸”或“XX 方”等。
A prescription (方剂) is a Traditional Chinese Medicine formula composed of multiple herbs, created based on principles of compatibility and clinical experience. The names usually follow the format:“Formula name + dosage form,” such as "XX Tang (汤)", "XX Wan (丸)", or "XX Fang (方)".

###Specification
1、遵循最小语义原则，在不影响文义的情况下，标注实体尽可能选取成词的最小单位。
Follow the principle of minimum semantic units: extract the smallest meaningful term without affecting overall meaning.
2、标注时医案中写明处方名的可直接标，如“二陈汤”“紫雪丹”“左金丸”。
If the formula name is explicitly mentioned in the medical case, extract it directly (e.g., Er Chen Tang, Zixue Dan, Zuo Jin Wan).
3、部分方剂标明方剂名称，剂型及序号的整体标注，如“叶熙钧治湿温方五”“叶熙钧治胃痛方三”。
When formulas include the physician’s name, indication, or numbering, extract the entire name as one entity (e.g., Ye Xijun’s Damp-Heat Formula No.5).

###Examples
input1：洪劳心营耗，风火交炽，饮啖酒肉，湿热内壅，络虚脉实，肉肿如痹。当此小满，阳气大泄，一阴未复，致内风挟阳上巅，耳目孔窍不清，舌苔黄浓，并不大渴，虽与客热不同，但口中酸浊吐痰。酒客不喜甘药，议进滋肾丸。
output1：['滋肾丸']
input2：程脉沉，喘咳浮肿，鼻窍黑，唇舌赤，渴饮则胀急，大便解而不爽。此秋风燥化，上伤肺气，气壅不降，水谷汤饮之湿痹阻经隧，最多坐不得卧之虑。法宜开通太阳之里，用仲景越婢、小青龙合方。若畏产后久虚，以补温暖，斯客气散漫，三焦皆累，闭塞告危矣。桂枝木、杏仁、生白芍、石膏、茯苓、炙草、干姜、五味。
output2：['仲景越婢','小青龙合方']
input3：顾，向年操持劳神，心阳动上亢，挟肝胆相火、肾中龙火自至。阴藏之火，直上巅顶，贯串诸窍。由情志内动而来，不比外受六淫客邪之变火，医药如凉药清肺不效，改投引火归源以治肾。诊脉坚而搏指，温下滋补，决不相投。仿东垣王善甫法，用滋肾丸。
output3：['滋肾丸']
input4：男，脉细而急，咳嗽痰不滑利。胸筑气促，曾经失血。阴虚肺热所致。叶熙钧治血溢方十七。料豆衣一钱五分，炙桑皮一钱五分，茯苓一钱五分，丹皮一钱，瓜蒌皮一钱五分，炒芩一钱，生苡仁三钱，马兜铃一钱，川贝母一钱五分，炙知母八分，紫菀一钱五分，枇杷叶一钱五分。
output4：['叶熙钧治血溢方十七']
input5：男，脾胃清阳失展，中州不司运化，气寒泛引作疼，怯冷口甘若糜，脉细舌白。譬之阴霾滔天，非离照当空，浊阴焉能退避。从温化法。叶熙钧治胃痛方五。制附片三分，黑炮姜六分，生茅术一钱，上川朴一钱，淡吴萸八分，元胡索一钱五分，小青皮八分，法夏一钱，炒香附一钱五分，广木香四分，公丁香四分。加毕澄茄八分，白蔻仁四分。
output5：['叶熙钧治胃痛方五']

###Error Analysis
1、实体抽取结果一般只输出一个方剂，医案中描述的病人前期误治、用之无效、用之症状加重的方剂不抽取，只抽取最终用于治疗的方剂。
Only extract the formula that was ultimately used for treatment. Do not extract formulas that were used earlier in the case but were ineffective, misprescribed, or worsened the patient’s condition.
2、注意有些医案中引用文献中的理论来解释病情，并未在实际治疗中使用，如：“仲景曰”、“《内经》曰”等。充分考虑上下文信息，注意区分引用文献中的方剂和实际应用中的方剂，请勿抽取引用文献中的方剂。
Do not extract formulas that are merely cited from classical sources (e.g., Shang Han Lun, Huangdi Neijing) to explain theory. These are not part of the actual treatment. Look for cues like “Zhang Zhongjing said” or “according to the Neijing.” Always assess context carefully.
3、注意区分中药与方剂。中药实体通常指单一的中药材，而方剂名称则是由多种中药组成的复合名称，勿将中药误认为方剂名称。
A formula (prescription) is composed of multiple herbs and has a recognized name. Do not mistake individual herbs for a prescription. Only extract compound formulas with proper names like Zuo Jin Wan or Er Chen Tang.
4、请充分考虑上下文信息，不抽取医案中方剂与中药并行描述的方剂实体。
If a formula is mentioned alongside a list of individual herbs, and it is unclear whether the formula was actually prescribed, do not extract it. Ensure that the formula is clearly used in the actual treatment context.

Fig. 1 Prompt Framework for Extracting TCM Prescriptions Using Large Language Models.

Table 3 Entity-level performance of the Prompt2 baseline for Traditional Chinese Medicine (TCM) entity recognition. This table reports Precision (P), Recall (R), and F1-score for each entity type under the Prompt2 setting, together with the macro-average across entity types, serving as the baseline for subsequent LLM-assisted and expert-refined comparisons.

Entity Type	**P**	**R**	**F1**
Symptoms	0.6528	0.6667	0.6095
Tongue Diagnosis	0.1667	0.1667	0.1667
Pulse Diagnosis	0.0833	0.1667	0.1111
Etiology & Pathogenesis	0.1905	0.2143	0.1667
Treatment Principles	0.3333	0.3333	0.3333
TCM Formulas	0.3472	0.8333	0.4833
TCM Herbs	0.5889	0.8803	0.6809
Average	0.3375	0.4659	0.3645

Note: Metrics are reported in decimal form (e.g., $0.5000 = 50.00\%$). Average is the unweighted mean over entity types.

Table 4 Entity-level performance of the Prompt2+LLM setting for TCM entity recognition. Compared with the Prompt2 baseline, this table summarizes how introducing an LLM-assisted step affects Precision (P), Recall (R), and F1-score for each entity type, and reports the macro-average to quantify overall gains or regressions after LLM assistance.

Entity Type	**P**	**R**	**F1**
Symptoms	0.4167	0.4444	0.4278
Tongue Diagnosis	0.1667	0.1667	0.1667
Pulse Diagnosis	0.5833	0.6667	0.6111
Etiology & Pathogenesis	0.3333	0.4286	0.3571
Treatment Principles	0.5000	0.5833	0.5278
TCM Formulas	0.5972	1.0000	0.7056
TCM Herbs	0.6583	0.9405	0.7545
Average	0.4651	0.6043	0.5072

Note: Metrics are reported in decimal form. Average is the unweighted mean over entity types, enabling fair comparison across settings.

Table 5 Entity-level performance after expert refinement based on LLM outputs for TCM entity recognition. This table reports Precision (P), Recall (R), and F1-score per entity type when domain experts review and refine the LLM-assisted results, reflecting the upper-bound quality attainable via human-in-the-loop correction; the macro-average summarizes the overall improvement level.

Entity Type	**P**	**R**	**F1**
Symptoms	0.6806	0.6667	0.6540
Tongue Diagnosis	0.3333	0.3333	0.3333
Pulse Diagnosis	0.7500	0.8333	0.7778
Treatment Principles	0.8333	0.7500	0.7778
Etiology & Pathogenesis	0.4762	0.5714	0.5000
TCM Formulas	1.0000	1.0000	1.0000
TCM Herbs	0.7972	0.9220	0.8336
Average	0.6958	0.7252	0.6966

Note: Metrics are reported in decimal form. Average is the unweighted mean over entity types.

Table 6 F1-score improvements (ΔF1) across settings for TCM entity recognition (recomputed). For each entity type, ΔF1(LLM–Prompt2) measures the gain from adding LLM assistance over the Prompt2 baseline; ΔF1(Expert–Prompt2) measures the gain of expert refinement over baseline; and ΔF1(Expert–LLM) isolates the additional benefit contributed by expert correction beyond the LLM-assisted output. Macro-average ΔF1 summarizes overall changes across entity types.

Entity Type	**ΔF1 (LLM–Prompt2)**	**ΔF1 (Expert–Prompt2)**	**ΔF1 (Expert–LLM)**
Symptoms	-0.1817	0.0444	0.2262
Tongue Diagnosis	0.0000	0.1667	0.1667
Pulse Diagnosis	0.5000	0.6667	0.1667
Treatment Principles	0.1944	0.4444	0.2500
Etiology & Pathogenesis	0.1905	0.3333	0.1429
TCM Formulas	0.2222	0.5167	0.2944
TCM Herbs	0.0737	0.1528	0.0791
Average	0.1427	0.3321	0.1894

Note: Positive ΔF1 indicates improvement, while negative values indicate degradation. All values are reported in decimal form.

Table 7 Performance Evaluation and Error Analysis of Baichuan4-Turbo for Traditional Chinese Medicine Entity Recognition

Entity Type	Metric	0-shot	1-shot	2-shot	3-shot	4-shot	5-shot	Error Analysis
Symptoms	Precision	49.45	52.35	53.79	56.54	59.53	59.41	62.33
	Recall	68.62	65.05	64.88	61.77	62.36	62.85	61.56
	F1-Score	54.94	56.00	56.76	56.84	58.84	59.07	60.11
Tongue Diagnosis	Precision	54.00	95.75	95.50	95.75	96.00	96.00	95.00
	Recall	54.00	96.00	96.00	96.00	96.00	96.00	95.00
	F1-Score	54.00	95.83	95.67	95.83	96.00	96.00	95.00
Pulse Diagnosis	Precision	66.71	57.67	62.92	67.17	76.79	72.21	81.00
	Recall	67.17	57.60	62.85	67.10	76.42	71.59	79.83
	F1-Score	66.80	57.54	62.79	67.04	76.47	71.80	80.24
Etiology & Pathogenesis	Precision	31.50	46.72	44.06	48.58	46.89	50.14	58.59
	Recall	51.56	62.53	61.19	63.24	60.86	62.93	60.04
	F1-Score	36.69	50.10	48.12	51.70	49.71	52.84	57.06
Treatment Principles	Precision	29.46	35.96	40.42	41.20	45.71	45.00	67.30
	Recall	34.67	38.00	41.58	42.50	47.42	45.50	67.75
	F1-Score	30.75	36.34	40.31	41.20	45.87	44.63	66.92
TCM Formulas	Precision	93.18	92.93	92.10	93.18	93.93	93.93	94.56
	Recall	97.50	97.75	96.75	97.50	98.25	98.50	97.50
	F1-Score	94.23	94.23	93.32	94.32	95.07	95.07	95.20
TCM Herbs	Precision	85.60	86.37	86.72	86.67	87.20	87.36	89.18
	Recall	89.91	90.06	90.52	90.31	90.32	91.03	90.97
	F1-Score	86.94	87.51	87.94	87.78	88.19	88.53	89.72
Average	Precision	58.56	66.82	67.93	69.87	72.29	72.01	78.28
	Recall	66.20	72.43	73.40	74.06	75.95	75.48	78.95
	F1-Score	60.62	68.22	69.27	70.67	72.88	72.56	77.75

Note: All values are reported as percentages. Best performance for each entity type is achieved through error analysis refinement.

Table 8 Performance Evaluation and Error Analysis of DeepSeek-V3 for Traditional Chinese Medicine Entity Recognition

Entity Type	Metric	0-shot	1-shot	2-shot	3-shot	4-shot	5-shot	Error Analysis
Symptoms	Precision	71.58	68.52	71.59	77.75	77.90	80.66	83.59
	Recall	78.36	75.64	78.00	82.05	78.99	80.19	81.12
	F1-Score	73.29	70.28	73.03	78.46	77.33	79.50	81.51
Tongue Diagnosis	Precision	94.00	97.50	96.50	97.50	97.50	97.00	97.25
	Recall	94.00	97.50	96.50	97.50	97.50	97.00	97.25
	F1-Score	85.38	81.67	82.67	83.17	87.00	88.00	93.75
Pulse Diagnosis	Precision	85.30	81.85	82.60	83.10	86.55	87.34	93.56
	Recall	85.17	81.63	82.46	82.96	86.63	87.49	93.40
	F1-Score	58.53	63.15	64.81	67.81	68.92	73.97	76.84
Etiology & Pathogenesis	Precision	71.02	68.16	67.56	73.74	72.50	74.63	73.18
	Recall	61.50	63.23	64.34	68.81	69.05	73.03	73.70
	F1-Score	57.69	56.42	56.99	67.73	70.20	74.32	83.48
Treatment Principles	Precision	63.67	58.75	59.67	70.50	73.42	76.92	83.25
	Recall	59.35	57.13	57.82	68.34	71.10	75.06	82.99
	F1-Score	92.43	91.35	91.60	91.56	92.28	93.61	97.25
TCM Formulas	Precision	93.84	92.80	93.75	96.75	98.25	98.25	96.33
	Recall	87.47	88.74	89.81	90.55	92.04	93.28	95.00
	F1-Score	91.59	91.36	92.27	93.07	94.55	94.93	95.27
TCM Herb	Precision	88.91	89.57	90.60	91.35	92.79	93.73	94.99
	Recall	78.15	78.19	79.14	82.26	83.69	85.83	89.60
	F1-Score	83.17	81.84	82.05	85.24	85.96	87.04	88.57
Average	Precision	79.44	78.88	79.71	82.88	84.03	85.79	88.64
	Recall	83.17	81.84	82.05	85.24	85.96	87.04	88.57
	F1-Score	79.44	78.88	79.71	82.88	84.03	85.79	88.64

Note: All values are reported as percentages. Best performance for each entity type is achieved through error analysis refinement.

Table 9 Performance Evaluation and Error Analysis of GLM-4-Plus for TCM Entity Recognition

Entity Type	Metric	0-shot	1-shot	2-shot	3-shot	4-shot	5-shot	Error Analysis
Symptoms	Precision	50.36	45.92	47.11	54.23	56.86	55.95	67.34
	Recall	71.84	71.19	69.16	72.44	71.91	69.52	70.66
	F1-Score	56.70	53.40	53.66	59.57	61.52	60.08	67.64
Tongue Diagnosis	Precision	93.50	96.50	97.50	95.50	97.50	97.00	96.50
	Recall	94.00	96.50	97.50	95.50	97.50	97.00	96.50
	F1-Score	76.42	68.33	74.58	76.58	82.25	80.50	88.21
Pulse Diagnosis	Precision	76.85	68.70	74.95	76.95	81.80	80.30	88.33
	Recall	76.54	68.42	74.67	76.67	81.88	80.29	88.18
	F1-Score	53.60	48.45	50.39	58.16	57.29	61.07	65.93
Etiology & Pathogenesis	Precision	45.86	40.51	43.40	52.15	51.94	56.94	64.63
	Recall	75.90	70.51	69.94	73.40	71.82	71.99	71.93
	F1-Score	45.61	37.77	39.14	40.48	42.59	42.55	58.81
Treatment Principles	Precision	50.00	42.58	42.00	42.50	44.75	45.00	59.58
	Recall	46.77	39.02	39.57	40.67	42.81	42.84	58.59
	F1-Score	85.73	92.11	91.58	92.33	92.36	92.83	95.44
TCM Formulas	Precision	87.11	93.42	93.14	93.64	93.54	94.06	95.74
	Recall	81.35	88.75	89.10	88.06	88.69	89.74	93.80
	F1-Score	89.59	90.53	90.73	89.43	91.25	91.64	94.06
TCM Herb	Precision	83.59	89.03	89.43	88.20	89.31	90.26	93.56
	Recall	68.40	67.13	68.92	71.33	73.17	73.64	80.68
	F1-Score	78.52	76.80	77.43	78.32	79.52	79.06	82.66
Average	Precision	71.14	69.75	71.19	73.20	74.84	75.08	80.88
	Recall	78.52	76.80	77.43	78.32	79.52	79.06	82.66
	F1-Score	71.14	69.75	71.19	73.20	74.84	75.08	80.88

Note: All values are reported as percentages. Best performance for each entity type is achieved through error analysis refinement.

Table 10 Performance Evaluation and Error Analysis of Qwen-Max for Traditional Chinese Medicine Entity Recognition

Entity Type	Metric	0-shot	1-shot	2-shot	3-shot	4-shot	5-shot	Error Analysis
Symptoms	Precision	54.30	52.78	55.05	59.50	65.91	66.66	71.87
	Recall	71.11	68.79	71.77	72.82	73.16	73.54	73.26
	F1-Score	59.40	57.32	59.68	62.85	67.54	68.38	71.35
Tongue Diagnosis	Precision	96.00	96.00	96.50	96.00	96.50	97.00	96.00
	Recall	96.00	96.00	96.50	96.00	96.50	97.00	96.00
	F1-Score	87.50	86.50	86.88	84.96	88.75	88.75	91.67
Pulse Diagnosis	Precision	87.09	86.05	86.37	84.27	87.79	87.79	90.99
	Recall	87.07	86.13	86.47	84.51	88.13	88.13	91.13
	F1-Score	50.59	50.85	52.07	50.89	53.83	58.61	63.55
Etiology & Pathogenesis	Precision	67.30	66.37	67.53	71.53	70.71	71.41	68.65
	Recall	55.03	54.78	55.96	56.45	58.33	61.88	64.51
	F1-Score	63.37	62.13	59.43	61.51	60.83	57.55	75.21
Treatment Principles	Precision	65.92	64.00	61.75	63.75	63.50	60.33	76.58
	Recall	64.04	62.39	59.94	61.97	61.39	58.22	75.44
	F1-Score	91.96	97.25	92.57	93.61	92.88	92.68	94.71
TCM Formulas	Precision	97.75	97.25	98.50	99.25	98.25	97.25	97.75
	Recall	93.91	94.93	94.19	93.88	95.53	95.74	93.76
	F1-Score	84.64	89.51	90.51	90.39	92.71	92.49	93.76
TCM Herbs	Precision	88.65	92.61	92.20	94.52	93.27	93.74	94.14
	Recall	85.77	90.44	90.80	93.28	92.65	93.74	93.74
	F1-Score	75.48	77.62	76.14	76.69	78.77	79.11	83.82
Average	Precision	81.97	81.58	82.09	82.80	83.49	82.94	85.34
	Recall	77.23	75.82	77.64	78.22	79.91	80.02	83.96
	F1-Score	77.23	75.82	77.64	78.22	79.91	80.02	83.96

Note: All values are reported as percentages. Best performance for each entity type is achieved through error analysis refinement.

Table 11 Examples of QA Error Cases

Natural-Language Question	Retrieved Context	Model-Generated Answer	Ground-Truth Answer
What symptoms may occur with constipation?	name1: Disease messages: Constipation name2: Symptom Manifestations: intestinal impediment, stomach pain, constipation, summerheat-warmth, wind, damp-warmth, cough, fright wind, hypochondriac pain, summerheat disease, other miscellaneous pediatric disorders, constraint disorder, bleeding, stomach pain, autumn dryness, vomiting, malaria, jaundice, liver wind, cold damage, sleep bruxism, wind-warmth, stroke	Constipation may present with symptoms such as intestinal impediment, stomach pain, constipation, summerheat-warmth, wind, damp-warmth, cough, fright wind, hypochondriac pain, summerheat disease, other miscellaneous pediatric disorders, constraint disorder, bleeding, stomach pain, autumn dryness, vomiting, malaria, jaundice, liver wind, cold damage, sleep bruxism, wind-warmth, and stroke.	According to *Xin'an Medical Cases*, constipation may present with symptoms such as cough, hypochondriac pain, bleeding, stomach pain, vomiting, sleep bruxism, wind, and liver wind. Please take the patient's individual condition into consideration and consult a qualified physician to determine the most appropriate treatment plan.
What symptoms may arise when there is lower deficiency with upper excess?	name1: Etiology and Pathogenesis messages: Lower deficiency with upper excess name2: Symptom Produces: constipation with blood in the stool, drooling, shortness of breath on exertion, dryness of the throat and tongue, vertigo and headache	Lower deficiency with upper excess may produce symptoms such as constipation with blood in the stool, drooling, dryness of the throat and tongue, and vertigo and headache.	According to *Xin'an Medical Cases*, lower deficiency with upper excess may produce symptoms such as constipation with blood in the stool, drooling, shortness of breath on exertion, dryness of the throat and tongue, and vertigo and headache. Please take the patient's individual condition into consideration and consult a qualified physician to determine the most appropriate treatment plan.