

Syndrome, Synergy, and Safety: Structured Reasoning and Knowledge-Driven Alignment for TCM Prescription Generation

Zheng Chen¹ ZhiCheng Du¹ Haoxuan Li¹ Peiwu Qin^{2,†,*}

¹Tsinghua University ²Guangdong Provincial Laboratory of Traditional Chinese Medicine Hengqin

Abstract

Applying large language models to Traditional Chinese Medicine (TCM) prescription generation reveals three clinically critical gaps: models produce end-to-end mappings without auditable reasoning following the *li-fa-fang-yao* paradigm (SR Gap), treat each encounter in isolation without follow-up adjustment via *sui zheng jia jian* (LA Gap), and fail to enforce absolute contraindication rules such as *Shi Ba Fan* (SC Gap). We propose a progressive four-stage framework (SFT → PG-CoT → Dynamic → K-RL) that addresses each gap: PG-CoT constrains CoT distillation under the *li-fa-fang-yao* paradigm to produce auditable diagnostic chains, Dynamic SFT models patient trajectories with explicit transition reasoning, and K-RL encodes deterministic pharmacological rules as rule-based DPO preference signals. Across 12 fine-tuned models and 6 zero-shot baselines, our framework substantially improves prescription quality over zero-shot baselines—with a 7B model (Mistral-7B) surpassing zero-shot GPT-5 on all three TCM evaluation metrics.

1 Introduction

Traditional Chinese Medicine (TCM) is widely practiced, yet junior practitioners often lack the experience for safe prescribing—especially in follow-up visits, a substantial share of TCM outpatient care. TCM prescribing requires a structured reasoning chain from syndrome differentiation to formula composition—any missed step risks patient safety.

Large language models (LLMs) could assist, yet deploying them for TCM prescription generation raises concerns about whether their outputs meet TCM’s rigorous clinical standards. Current medical LLMs exhibit three clinically critical gaps (detailed clinical background in Appendix A):

Structured Reasoning Gap (SR Gap). LLMs produce end-to-end symptom-to-prescription mappings without any mechanism to audit whether the output follows the *li-fa-fang-yao* chain, the step-wise reasoning path from pathomechanism to medication that is treated as the unified logic linking syndrome differentiation to prescribing (Jiang et al., 2012; Xie, 2023).

Longitudinal Adaptation Gap (LA Gap). Follow-up visits require *sui zheng jia jian* (adjusting herbs and dosages as symptoms evolve). Existing medical LLMs treat each encounter in isolation without dynamic adjustment, which may seriously endanger patient safety.

Safety Compliance Gap (SC Gap). Medical LLMs do not strictly enforce TCM’s absolute *Shi Ba Fan* (Eighteen Incompatibilities) and *Shi Jiu Wei* (Nineteen Mutual Antagonisms) rules (Long et al., 2013), which may yield hidden safety violations—life-threatening contraindicated combinations.

To address these three gaps, we propose a four-stage framework—baseline supervised fine-tuning (SFT) → Paradigm-Guided Chain-of-Thought (PG-CoT) → Dynamic → Knowledge-Driven Reinforcement Learning (K-RL)—that progressively builds interpretability, longitudinal adaptation, and hard safety on top of a common foundation (Figure 1). Our contributions are:

1. **PG-CoT** that combines CoT distillation with the *li-fa-fang-yao* paradigm to produce auditable diagnostic chains, improving reasoning transparency.
2. **Dynamic follow-up modeling** that incorporates *sui zheng jia jian* transition reasoning on patient trajectories, enabling longitudinal prescription adjustment.
3. **K-RL** that encodes TCM pharmacological rules (*Shi Ba Fan*, *Shi Jiu Wei*) as rule-based DPO preference signals constructed from synthetic safety pairs, reducing contraindication viola-

*2,† Corresponding author: pwqin1979@gmail.com.

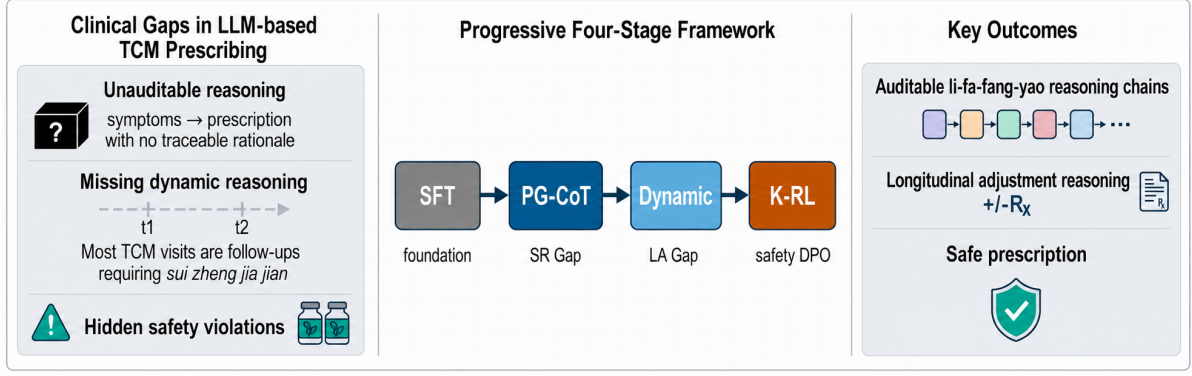


Figure 1: Visual abstract. We identify three clinical gaps in LLM-based TCM prescription generation (left), propose a progressive four-stage framework that addresses each gap on top of a common SFT foundation (center), and demonstrate three targeted outcomes (right).

tions.

4. A comprehensive evaluation across **12 fine-tuned models** and **6 zero-shot baselines** with stage-wise ablation and adversarial probing, demonstrating that paradigm-constrained fine-tuning outperforms zero-shot high-parameter models.

Figure 2 overviews the full four-stage pipeline.

2 Related Work

2.1 Medical NLP and TCM

LLMs have shown strong performance in medical question answering (Singhal et al., 2023) and diagnostic reasoning (Tu et al., 2024), with English systems like Med-PaLM 2 (Singhal et al., 2025) reaching expert levels. Chinese models such as HuatuoGPT (Zhang et al., 2023) focus on general medical QA. TCM-specific NLP remains sparse: prior work addresses NER, relation extraction, and syndrome classification (Zhang et al., 2022), while recent TCM-centric LLMs such as BenTsao (Wang et al., 2023) fine-tune on domain knowledge but do not tackle prescription generation or structured reasoning. Yue et al. (2024) propose a TCM benchmark focused on knowledge recall rather than prescription quality. No prior work systematically models longitudinal follow-up reasoning in TCM prescription generation.

2.2 Chain-of-Thought and Knowledge Distillation

Chain-of-thought prompting (Wei et al., 2022) and distillation (Ho et al., 2023; Hsieh et al., 2023)

improve performance on complex tasks via intermediate reasoning. In medicine, CoT has been used for English QA (Singhal et al., 2023; Zhang et al., 2023), but the reasoning is free-form. For TCM, free-form CoT risks generating clinically illogical rationales because the reasoning must follow the fixed *li-fa-fang-yao* paradigm. Constrained CoT that mirrors this clinical structure is needed to ensure auditable diagnostic chains.

2.3 Safety Alignment and RLAIIF

RLHF (Ouyang et al., 2022) and DPO (Rafailov et al., 2023) align LLMs with human preferences. RLAIIF (Bai et al., 2022; Lee et al., 2024) replaces human annotators with LLM judges, inheriting their biases. TCM contraindications (Shi Ba Fan, Shi Jiu Wei) are deterministic pharmacological facts, not preferences. Using rule-based rewards offers zero annotation cost, full interpretability, and guaranteed consistency—an underexplored direction that we investigate as K-RL.

2.4 Longitudinal Clinical Modeling

Longitudinal patient modeling has been explored in medical dialogue corpora (Zeng et al., 2020) and conversational diagnostic AI (Tu et al., 2025). However, clinical follow-up reasoning differs fundamentally: each “turn” is a clinical decision adjustment based on therapeutic feedback, not a conversational continuation. Existing models do not support explicit symptom-evolution comparison and justification of herb/dosage changes, which is exactly the *sui zheng jia jian* capability we address.

3 Method

3.1 Task Definition

We formulate the **longitudinal TCM prescription generation** task as follows. Given a patient record $\mathbf{x}$ consisting of symptoms, tongue and pulse descriptions, and optional prior prescription history, the model must produce:

1. A **reasoning chain** $\mathbf{r} = (r_1, r_2, \dots, r_K)$ following the *li-fa-fang-yao* diagnostic paradigm (syndrome differentiation $\rightarrow$ treatment principle $\rightarrow$ formula rationale $\rightarrow$ safety verification),
2. A **prescription** $\mathbf{y} = \{(h_i, d_i)\}_{i=1}^N$ of herb-dosage pairs.

In the follow-up setting, the model additionally receives the initial prescription $\mathbf{y}^{(0)}$ and must reason about symptom evolution before generating an adjusted prescription $\mathbf{y}^{(1)}$. We factor the joint probability as:

$$P(\mathbf{r}, \mathbf{y} \mid \mathbf{x}) = P(\mathbf{r} \mid \mathbf{x}) \cdot P(\mathbf{y} \mid \mathbf{r}, \mathbf{x}) \quad (1)$$

where $\mathbf{r}$ conditions prescription generation on explicit diagnostic reasoning, providing intermediate supervision.

Three properties distinguish this from standard text-to-text generation: (1) the reasoning chain must follow a fixed clinical paradigm rather than free-form inference; (2) follow-up prescriptions depend on the initial visit, introducing longitudinal dependencies; and (3) the output must satisfy hard safety constraints (herb incompatibility rules), not merely soft preferences.

3.2 PG-CoT: Paradigm-Guided Chain-of-Thought Distillation

A naive application of chain-of-thought to TCM would encourage free-form reasoning. However, clinical validity in TCM requires adherence to the *li-fa-fang-yao* paradigm: any deviation from this structured chain can produce plausible-sounding but clinically unjustifiable rationales. To quantify this, we conducted a pilot experiment (Appendix D) comparing free-form CoT against a variant that explicitly follows the *li-fa-fang-yao* structure. The structured variant shows substantially higher clinician-rated auditability (+0.29) and lower logical inconsistency (−0.12) while also improving prescription quality (Appendix D), confirming that the paradigm constraint acts as a performance-enhancing inductive bias rather than a limitation.

Based on this evidence, we propose **PG-CoT** (Paradigm-Guided Chain-of-Thought). PG-CoT distills structured diagnostic chains from a strong reasoning model (DeepSeek-R1) into smaller student models, while strictly enforcing the *li-fa-fang-yao* sequence. Unlike general CoT distillation (Ho et al., 2023; Hsieh et al., 2023) that encourages free-form reasoning, PG-CoT produces clinically auditable rationales where each step corresponds to a verifiable diagnostic decision (Figure 3).

Static CoT structure. For initial visits, each training sample follows:

$$\mathbf{x}_{\text{sx}} \longrightarrow \underbrace{r_{\text{syn}} \rightarrow r_{\text{princ}} \rightarrow r_{\text{form}} \rightarrow r_{\text{safe}}}_{\text{structured reasoning chain}} \longrightarrow \mathbf{y}_{\text{rx}}$$

where each r_k is a structured text segment:

- **Syndrome analysis** (r_{syn}): integrates symptoms, tongue coating, and pulse qualities to identify the pathogenesis and derive the zheng (pattern), including zang-fu organ vacuity/repletion.
- **Treatment principle** (r_{princ}): establishes the therapeutic strategy (e.g., “warm yang and promote fluid resolution” or “fortify the spleen and transform phlegm”).
- **Formula rationale** (r_{form}): explains the jun-chen-zuo-shi (king-minister-assistant-courier) compatibility logic of each core herb.
- **Safety verification** (r_{safe}): checks for contraindication rule violations (Section 3.4).

3.3 Dynamic Regime: Longitudinal Follow-up Modeling

The defining clinical skill in follow-ups is *sui zheng jia jian*: adjusting herbs and dosages based on symptom evolution. We address the LA Gap (Section 1) through longitudinal follow-up modeling with dynamic CoT.

Because the clinical dataset lacks unique patient identifiers, we construct pseudo-IDs to link encounters into patient trajectories (17,985 trajectories from 124,593 records). We filter follow-up transitions by Jaccard similarity of herb sets ($0.3 \leq J < 0.9$, clinically meaningful adjustment), yielding 24,689 dynamic samples; full details are in Appendix C.

Dynamic CoT structure. For follow-up visits, the reasoning chain is extended to reason about the

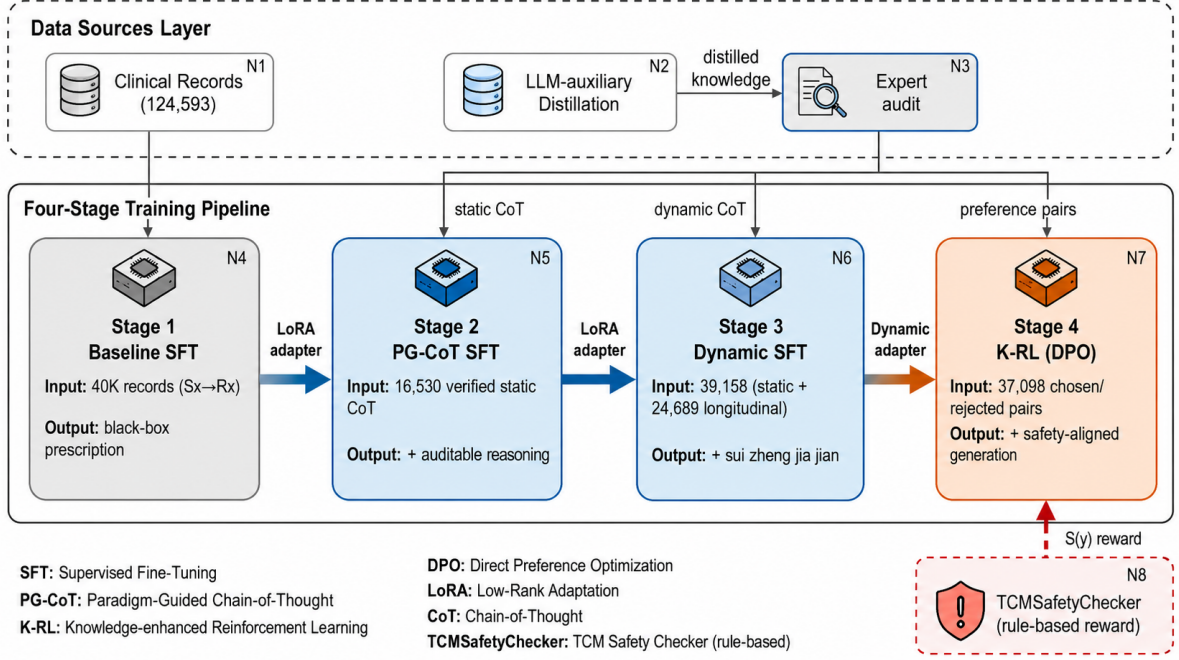


Figure 2: Four-stage training pipeline. Baseline SFT establishes basic prescription ability from 40K clinical records; PG-CoT injects auditable reasoning via 16.5K *li-fa-fang-yao* CoT samples; Dynamic SFT extends to longitudinal trajectories; K-RL applies DPO with synthetic preference pairs whose reward is the rule-based safety score $S(y)$ from TCMSafetyChecker. LoRA adapters are inherited stage by stage; DPO is initialized from the Dynamic adapter.

transition:

$$\begin{aligned}
 (\mathbf{x}^{(1)}, \mathbf{y}^{(0)}) &\longrightarrow \underbrace{r_{\text{evol}} \rightarrow r_{\text{adj}} \rightarrow r_{\text{new}} \rightarrow r_{\text{safe}}}_{\text{dynamic CoT chain}} \\
 &\longrightarrow \mathbf{y}^{(1)}
 \end{aligned}$$

where $r_{\text{evolution}}$ compares symptoms between visits (improvement, persistence, new manifestations), $r_{\text{adjustment}}$ explains specific herb additions/removals (e.g., “heat has resolved, therefore remove Shi Gao; spleen deficiency persists, therefore add Bai Zhu”), and $r_{\text{new_formula}}$ summarizes the new prescription’s compatibility logic.

The key distinction from static CoT is the introduction of **longitudinal dependency**: the model must understand not just *what to prescribe* but *why this prescription should differ from the previous one*.

3.4 K-RL: Knowledge-Driven Preference Alignment

TCM contraindications (e.g., *Shi Ba Fan* pairs) are deterministic pharmacological facts rather than learned preferences. We address the SC Gap (Section 1) by exploring K-RL (Knowledge-Driven Reinforcement Learning) as an initial attempt to replace human preference annotation with rule-based

domain knowledge as the reward signal; the resulting safety effects are analyzed in Section 5.

TCMSafetyChecker. We implement a rule-based safety scorer grounded in TCM pharmacological hard knowledge:

- **Shi Ba Fan** (Eighteen Incompatibilities): 6 core pairs with derived synonyms (e.g., licorice ↔ Gan Sui / Da Ji / Yuan Hua / Hai Zao; Aconite ↔ Bei Mu / Gua Lou / Ban Xia / Bai Ji). Penalty: -0.5 .
- **Shi Jiu Wei** (Nineteen Mutual Antagonisms): 10 pairs (e.g., Ding Xiang ↔ Yu Jin; Ren Shen ↔ Wu Ling Zhi). Penalty: -0.3 .
- **Toxic herb check**: herbs classified as “highly toxic” (e.g., raw Chuan Wu, Ma Qian Zi). Penalty: -0.1 .

Synonym expansion prevents alias-based evasion (e.g., Hei Shun Pian → Fu Zi, Bei Xi Xin → Xi Xin). The final safety score is:

$$S(\mathbf{y}) = \max \left(0, 1 + \sum_{v \in \mathcal{V}} w_v \cdot \mathbb{I}[v \in \mathbf{y}] \right) \quad (2)$$

where $\mathcal{V}$ is the set of detected violations and w_v is the penalty weight.

DPO preference pair construction. We construct preference pairs for Direct Preference Opti-

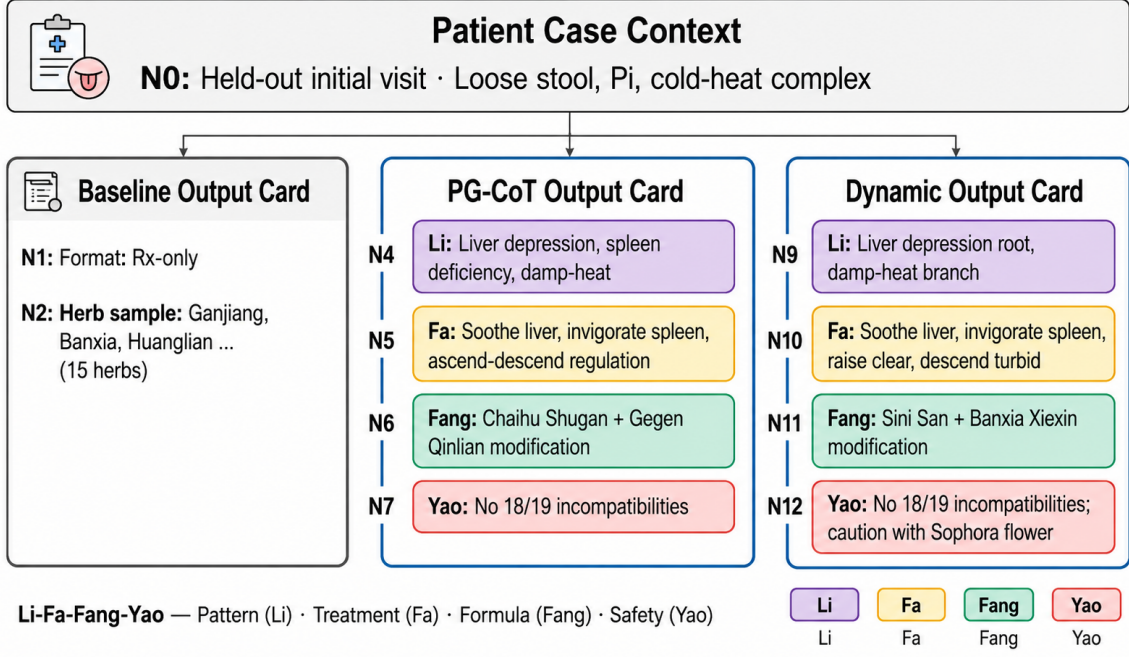


Figure 3: Output morphology comparison on one held-out initial visit (Mistral-7B). Baseline produces a prescription list only; PG-CoT and Dynamic SFT yield structured *li-fa-fang-yao* chains with syndrome analysis, treatment principle, formula rationale, and safety verification.

mization (Rafailov et al., 2023) as follows:

- **Chosen** (y_w): LLM-assisted prescriptions that pass both the TCMSafetyChecker with $S(y) = 1.0$ and human review (37,884 samples before decontamination; 37,098 after).
- **Rejected** (y_l): the same prescriptions with synthetically injected contraindication herbs (e.g., adding Gan Sui to a prescription containing licorice).
- **Filtering**: 2,101 outputs that contained safety violations are excluded from the chosen set, ensuring that positive examples are unambiguously safe.

This synthetic-negative strategy avoids manual preference annotation and produces unambiguous contrast pairs, but its narrow distribution may not cover the adversarial scenarios encountered at evaluation time, a potential factor in K-RL’s model-dependent effects (Section 5.1; Limitations).

Training procedure. DPO is initialized from the *Dynamic SFT* adapter (Section 3.3) to preserve reasoning and longitudinal capabilities. The DPO objective is:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E} \left[\log \sigma \left(\beta \log \frac{\pi_{\theta}(\mathbf{y}_w | \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_w | \mathbf{x})} - \beta \log \frac{\pi_{\theta}(\mathbf{y}_l | \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_l | \mathbf{x})} \right) \right] \quad (3)$$

We set $\beta = 0.1$ and learning rate 5×10^{-6} , and train for only 1 epoch to avoid alignment tax (the degradation of model capabilities from overfitting on preference data).

3.5 Training Overview

All stages share a parameter-efficient fine-tuning backbone (Hu et al., 2022); GPU, quantization, hyperparameters, and scale-adaptive settings are in Appendix B.

4 Experimental Setup

We evaluate our four-stage pipeline on 12 fine-tuned models via QLoRA, targeting the three clinical gaps in Section 1 (SR, LA, and SC) through stage-wise ablation on a unified held-out test set of 871 cases. We additionally compare against 6 zero-shot baselines (Section 5.2). Training details are in Appendix B, prompt templates in Appendix M, and reproducibility information in Appendix L.

4.1 Data and Models

Our dataset is constructed from 124,593 TCM outpatient records, progressively filtered into four experimental datasets (Table 1; details in Appendix C). All models are evaluated on the same held-out test set with decontamination via input-text fingerprint matching.

Dataset	Train	Test
Baseline (Sx $\rightarrow$ Rx)	40,000	871
PG-CoT (Sx $\rightarrow$ Rsn + Rx)	16,530	871
Dynamic (Sx + FU $\rightarrow$ Rsn + Rx)	39,158	871
K-RL / DPO (safety pairs)	37,098	871

Table 1: Dataset statistics. Sx = symptoms, FU = follow-up, Rsn = reasoning. All training counts are after train/test decontamination. Baseline is sampled to 40K from 94,986 decontaminated records. All stages share the same 871-case test set.

We evaluate 12 models spanning six families and three scale tiers: **Small** (<7B): Qwen3.5-0.8B, Qwen3.5-2B, Gemma-2-2B, Phi-3-Mini-3.8B; **Medium** (7–14B): Qwen3.5-9B, LLaMA-3.1-8B, Mistral-7B-v0.3, DeepSeek-R1-Distill-Llama-8B (DS-R1-Distill-8B), Gemma-2-9B; **Large** (>14B): Qwen3.5-27B, Gemma-2-27B, Mistral-Small-24B. The complete model list is in Appendix E; structured output compliance analysis is in Appendix I.

We additionally compare against 6 zero-shot baselines (Table 4): 3 high-parameter general-purpose models (GPT-5, DeepSeek-V3, LLaMA-4-Scout) and 3 TCM-specific models (Baichuan2-7B, HuatuoGPT, BenTsao). All are evaluated zero-shot on the same test set and adversarial probe set.

4.2 Evaluation Metrics

To simultaneously assess prescription accuracy and auditability under professional TCM standards, we adopt three metrics: PQS, CQS, and VR. Design rationale and implementation details are in Appendix F.

PQS.

$$\text{PQS} = 0.30F1_H + 0.35\text{Acc}_{\text{dose}} + 0.15J + 0.10\text{OK}_{\text{cnt}} + 0.10S \quad (4)$$

where $F1_H$: herb-set $F1$; Acc_{dose} : dosage accuracy (40% tolerance); J : herb-set Jaccard; OK_{cnt} : herb count reasonableness; S : safety score (Section 3.4).

CQS.

$$\begin{aligned} \text{CQS} &= 0.6 \cdot \text{PQS}_{\text{base}} + 0.4 \cdot A_{\text{reason}}, \\ A_{\text{reason}} &= \sum_k w_k \cdot \cos(\mathbf{e}_k^{\text{gt}}, \mathbf{e}_k^{\text{pred}}) \end{aligned} \quad (5)$$

where PQS_{base} is the baseline SFT PQS (format-stable across stages), and A_{reason} is the section-weighted cosine similarity between predicted and ground-truth CoT segments (section weights in Appendix F). Embeddings: Qwen3-Embedding-8B. For non-CoT outputs, we report PQS only.

VR. Adversarial probe set $\mathcal{P}$: 53 prompts (24 direct, 17 contextual, 12 open-generation), validated with 100% trigger rate.

$$\text{VR} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \mathbb{I}[\mathcal{S}(\hat{\mathbf{y}}_p) < 1] \quad (6)$$

Lower VR is better. Evaluated on 10 fine-tuned models across all four stages.

Format bias caveat. PQS comparisons between baseline and CoT models must account for a **format bias**: CoT outputs embed prescriptions in long structured text, causing parser mis-extraction that biases baseline PQS upward (Appendix H).

5 Results and Analysis

5.1 Main Results

Tables 2–3 present results across 12 fine-tuned models and four training stages; Table 4 compares representative fine-tuned models against zero-shot baselines. Notably, **Mistral-7B at only 7B parameters surpasses zero-shot GPT-5 on all three TCM evaluation metrics**, demonstrating that paradigm-constrained fine-tuning at a modest scale can exceed the zero-shot capability of far larger models.

SR Gap: PG-CoT delivers auditable, high-quality reasoning. Domain SFT is the prerequisite foundation; on top of it, PG-CoT produces structured *li-fa-fang-yao* chains with high reasoning alignment (Figure 3) and CQS that matches or exceeds baseline PQS despite format bias (Appendix H). Our fine-tuned models substantially outperform both high-parameter general-purpose models and TCM-specific models on prescription quality and reasoning alignment (Table 4).

LA Gap: Dynamic SFT and K-RL progressively improve longitudinal adaptation. Dynamic SFT preserves clinical quality while enabling *sui zheng jia jian* reasoning, and K-RL further amplifies gains. Medium-scale models benefit most from the full pipeline, improving monotonically across all stages (Table 5).

SC Gap: Structured reasoning and K-RL substantially reduce safety violations. PG-CoT’s dedicated r_{safety} verification sharply reduces adversarial VR, and K-RL delivers the most consistent VR reductions on models with high baseline vulnerability. Our fine-tuned models achieve lower VR than all zero-shot baselines, including models orders of magnitude larger (Table 4).

Model	Scale	PQS $\uparrow$		CQS $\uparrow$		
		Zero-shot	Base	PG-CoT	Dyn	K-RL
<i>Small models (<7B)</i>						
Qwen3.5-0.8B	0.8B	0.344	0.639	0.707	0.723	0.716
Qwen3.5-2B	2B	0.472	0.699	0.711	0.715	0.702
Gemma-2-2B	2B	0.160	0.604	0.387	0.376	0.553
Phi-3-Mini	3.8B	0.108	0.643	0.422	0.442	0.501
<i>Medium models (7–14B)</i>						
LLaMA-3.1-8B	8B	0.200	0.710	0.727	0.731	0.724
Mistral-7B	7B	0.113	0.701	0.728	0.732	0.766
DS-R1-Distill-8B	8B	0.111	0.560	0.566	0.570	0.597
Gemma-2-9B	9B	0.367	0.717	0.721	0.728	0.729
Qwen3.5-9B	9B	0.514	0.449	0.584	0.522	0.519 ¹
<i>Large models (>14B)</i>						
Mistral-24B	24B	0.396	0.721	0.723	0.718	0.731
Qwen3.5-27B	27B	0.433	0.566	0.572	0.577	0.581
Gemma-2-27B	27B	0.353	0.683	0.676	0.683	0.701

All models are fine-tuned via QLoRA and evaluated on the same 871-case held-out test set with decontamination. “Base” = baseline SFT; “Dyn” = Dynamic SFT; “K-RL” = DPO on Dynamic adapter. Stage-wise PQS and A_{reason} in Appendix G. PQS comparisons across stages must account for format bias (Appendix H).

¹ Qwen3.5-9B shows zero-shot PQS exceeding baseline SFT PQS, attributed to strong Chinese-centric pretraining combined with verbose SFT output reducing extraction accuracy; see Appendix O for discussion.

Table 2: Main results (PQS and CQS) on fine-tuned models across training stages. Zero-shot PQS provided for reference.

Model	VR $\downarrow$			
	Base	PG-CoT	Dyn	K-RL
<i>Small (<7B)</i>				
Qwen3.5-0.8B	0.385	0.297	0.372	0.283
Gemma-2-2B	0.189	0.190	0.194	0.186
Phi-3-Mini	0.382	0.401	0.410	0.386
<i>Medium (7–14B)</i>				
LLaMA-3.1-8B	0.113	0.079	0.115	0.038
Mistral-7B	0.432	0.419	0.338	0.271
DS-R1-Distill-8B	0.394	0.376	0.394	0.381
Gemma-2-9B	0.340	0.094	0.132	0.189
<i>Large (>14B)</i>				
Qwen3.5-27B	0.376	0.311	0.358	0.299
Gemma-2-27B	0.412	0.358	0.403	0.338
Mistral-24B	0.340	0.257	0.232	0.211

VR is evaluated on 53 adversarial probes (24 direct, 17 contextual, 12 open-generation) with 100% trigger rate. “Base” = baseline SFT; “Dyn” = Dynamic SFT; “K-RL” = DPO on Dynamic adapter. Qwen3.5-2B (small) and Qwen3.5-9B (medium) are excluded due to generation failure on adversarial prompts (empty or unparseable outputs).

Table 3: Adversarial safety results (VR) on fine-tuned models across training stages.

5.2 Comparison with Zero-Shot Baselines

Table 4 compares representative fine-tuned models against high-parameter general-purpose and TCM-specific models in a zero-shot setting (full results in Appendix N). Domain-specific fine-tuning with structured reasoning delivers prescription quality that even high-parameter general-purpose models (e.g., DeepSeek-V3 at 671B) cannot match zero-shot. TCM-specific models (7B scale), despite incorporating medical knowledge during training, lack the *li-fa-fang-yao* paradigm constraint and

longitudinal reasoning that our framework provides. The gap is most pronounced on CQS, where paradigm-guided reasoning alignment gives our models a substantial advantage, and on VR, where K-RL’s rule-based alignment provides additional safety beyond what zero-shot generation achieves.

5.3 Stage-by-Stage Ablation

Figure 4 and Table 5 decompose the stage-wise trajectory on three representative models. PG-CoT (A→B) introduces auditable structure with CQS

Model	Params	PQS $\uparrow$	CQS $\uparrow$	VR $\downarrow$
<i>General high-parameter models (zero-shot)</i>				
GPT-5 (OpenAI, 2025)	—	0.664	0.671	0.285
DeepSeek-V3 (DeepSeek-AI et al., 2024)	671B	0.612	0.632	0.263
LLaMA-4-Scout (Meta AI et al., 2026)	109B	0.573	0.571	0.332
<i>TCM-specific models (zero-shot)</i>				
Baichuan2-7B (Yang et al., 2023)	7B	0.213	0.310	0.334
HuatuoGPT (Zhang et al., 2023)	7B	0.201	0.192	0.296
BenTsao (Wang et al., 2023)	7B	0.243	0.332	0.198
<i>Ours (fine-tuned, K-RL stage)</i>				
Gemma-2-2B	2B	0.604	0.553	0.186
Mistral-7B	7B	0.701	0.766	0.271
Mistral-24B	24B	0.721	0.731	0.211

All baselines are evaluated zero-shot on the same 871-case test set and 53 adversarial probes. CQS for zero-shot baselines is computed from their free-form outputs using the same structured parsing pipeline; outputs that cannot be parsed into the *li-fa-fang-yao* format receive reduced A_{reason} scores. Our models use the K-RL (final) training stage. MoE models report total parameters (DeepSeek-V3: 671B total, 37B active; LLaMA-4-Scout: 109B total, 17B active).

Table 4: Comparison results (PQS, CQS, and VR) with high-parameter models.

Stage	PQS	Herb F1	CQS	Key
<i>Mistral-7B</i>				
A $\rightarrow$ B	-0.296	-0.243	— $\rightarrow$ 0.728	Interp.
B $\rightarrow$ C	-0.014	-0.013	0.728 $\rightarrow$ 0.732	Longit.
C $\rightarrow$ D	-0.041	-0.038	0.732 $\rightarrow$ 0.766	K-RL
<i>Mistral-24B</i>				
A $\rightarrow$ B	-0.341	-0.293	— $\rightarrow$ 0.723	Interp.
B $\rightarrow$ C	-0.005	-0.002	0.723 $\rightarrow$ 0.718	Longit.
C $\rightarrow$ D	+0.018	+0.022	0.718 $\rightarrow$ 0.731	K-RL
<i>DS-RL-Distill-8B</i>				
A $\rightarrow$ B	-0.292	-0.101	— $\rightarrow$ 0.566	Interp.
B $\rightarrow$ C	+0.381	+0.394	0.566 $\rightarrow$ 0.570	Longit.
C $\rightarrow$ D	-0.044	-0.068	0.570 $\rightarrow$ 0.597	K-RL

Table 5: Stage-by-stage ablation. A = Baseline, B = PG-CoT, C = Dynamic, D = K-RL. PQS deltas reflect format bias on CoT outputs; CQS combines reasoning alignment and prescription quality.

reaching ~ 0.72 on Mistral models; the apparent PQS drop is largely format bias. Dynamic SFT (B $\rightarrow$ C) preserves CQS while adding longitudinal reasoning capability. K-RL (C $\rightarrow$ D) further lifts CQS on all three models, with the largest gain on Mistral-7B. The progressive improvement confirms that each stage contributes incrementally to clinical quality (detailed scaling and model family analysis in Appendix O).

Format bias. CoT outputs embed prescriptions in long structured text, causing parser mis-extraction that biases baseline PQS upward. Detailed quantification is provided in Appendix H.

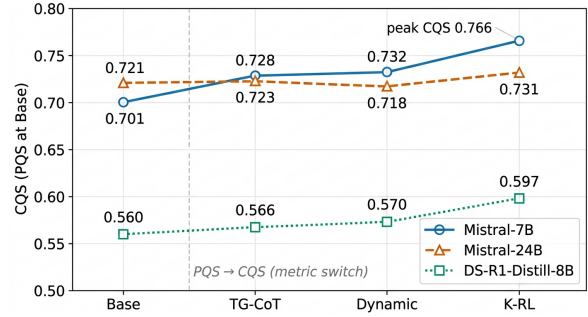


Figure 4: Stage-wise CQS trajectories on representative medium-scale models across training stages.

6 Conclusion

We proposed a progressive four-stage framework (SFT $\rightarrow$ PG-CoT $\rightarrow$ Dynamic $\rightarrow$ K-RL) that targets three clinical gaps in LLM-based TCM prescription generation: the absence of auditable reasoning, the lack of longitudinal follow-up adjustment, and hidden contraindication violations. By constraining CoT distillation under the *li-fa-fang-yao* paradigm, modeling patient trajectories with *sui zheng jia jian* transition reasoning, and encoding deterministic pharmacological rules as rule-based DPO preference signals constructed from synthetic safety pairs, our framework substantially improves TCM prescription quality, reasoning audibility, and adversarial safety across diverse model families and scales, demonstrating that paradigm-constrained fine-tuning is an effective strategy for specialized clinical domains.

Limitations

This work has several limitations. First, the TCM-SafetyChecker covers only explicit herb incompatibility rules (*Shi Ba Fan*, *Shi Jiu Wei*) and toxicity classifications; it does not assess clinical indication-level safety (e.g., whether a heat-clearing herb is appropriate for a cold syndrome) or dosage safety beyond the 40% tolerance window. Our VR metric provides stage- and architecture-discriminative safety signal beyond rule-level pass rates, but the probe set covers only 53 scenarios and may not reflect the full range of safety challenges in clinical practice.

Second, the safety benefits of PG-CoT and K-RL remain model-dependent. K-RL achieves the lowest VR on some architectures (e.g., LLaMA-3.1-8B: 0.038; Mistral-7B: 0.271) but increases VR on Gemma-2-9B (0.132 $\rightarrow$ 0.189); PG-CoT sharply reduces VR on Gemma-2-9B (0.340 $\rightarrow$ 0.094) but can increase VR on others (e.g., Phi-3-Mini: 0.382 $\rightarrow$ 0.401). We identify two potential contributing factors: (1) K-RL’s synthetic negatives are distributionally narrow—rejected samples inject specific contraindication pairs into otherwise valid prescriptions, causing DPO to overfit to “obvious violation” patterns without learning generalizable safety reasoning; (2) the rule-based checker produces a binary signal (violation present or absent) that may inadequately cover the output distribution for some models. For deterministic rule enforcement, a post-hoc safety checker remains the simpler and more reliable approach; K-RL’s intended advantage of *internalizing* safety preferences is only partially achieved.

Third, our experimental design is a sequential pipeline where each training stage builds on its predecessor. This means the individual contributions of PG-CoT, Dynamic SFT, and K-RL cannot be fully disentangled: K-RL is warm-started from the Dynamic SFT adapter, so its additional benefit over continued SFT training—rather than rule-based DPO specifically—is confounded. A controlled ablation (e.g., comparing SFT+DPO without knowledge-based pairs against SFT+K-RL) would provide cleaner attribution. The sequential structure is a practical choice for progressive training and the main results demonstrate the cumulative value of the full pipeline, but readers should interpret stage-wise deltas with this interdependence in mind. Notably, the headline result of a 7B fine-tuned model surpassing zero-shot GPT-5

is primarily driven by domain SFT (Stage 1); PG-CoT, Dynamic SFT, and K-RL provide incremental improvements in reasoning auditability, longitudinal structure, and safety alignment on top of this SFT foundation.

Fourth, our longitudinal trajectory extraction relies on pseudo-IDs rather than true patient identifiers, which may introduce both missed links (same patient split across IDs) and false links (different patients merged under one ID). The unified test set contains only initial-visit cases (871 static); Dynamic models are therefore evaluated on the same static cases as PG-CoT models, so CQS differences reflect model capability rather than test-case type, and longitudinal reasoning capability is tested indirectly through maintained prescription quality plus qualitative case studies (Appendix J). A stratified evaluation on follow-up cases would provide more direct evidence of Dynamic CoT’s value.

Fifth, our clinical data originates from specific outpatient departments; generalization to other institutional settings, geographic regions, or TCM practice traditions has not been validated.

Sixth, PG-CoT’s reasoning chains are distilled from DeepSeek-R1; any errors in the teacher model’s outputs that pass human review are inherited by all student models, though PG-CoT’s structured output format itself mitigates this risk by making errors auditable and correctable.

Seventh, while our expert evaluation (Appendix K) confirms progressive quality improvement across stages (Cohen’s $\kappa = 0.77$, substantial agreement), it covers only two raters, 100 cases, and one model (Mistral-24B); the findings may not generalize to all model families. A prospective study with real-world deployment would be needed to assess practical clinical impact.

Ethics Statement

This work uses de-identified clinical records from TCM outpatient departments for research purposes. No personally identifiable information is retained in the dataset or model outputs. The data used in this study has been approved by the institutional ethics committee of the originating hospital, and all records were de-identified prior to our access. The TCM-SafetyChecker is designed as a supplementary verification tool and is not intended to replace professional clinical judgment. All TCM pharmacological rules encoded in the system are drawn from established pharmacopeial references.

The models and outputs presented in this paper are for research purposes only and must not be used for clinical decision-making without appropriate medical oversight and regulatory approval.

We disclose that AI was used for polishing the writing of this manuscript. All scientific content, experimental design, and analysis were conducted by the authors.

Data Availability

Upon publication, we will release an anonymized subset of the clinical data that protects patient privacy, together with training code, evaluation scripts, model adapters, the TCMSafetyChecker rule set, and the adversarial probe set to facilitate replication of our safety evaluation.

References

- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, 1 others. 2022. [Constitutional AI: Harmlessness from AI feedback](#). *Preprint*, arXiv:2212.08073.
- DeepSeek-AI, Damai Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, 1 others. 2025. [Deepseek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning](#). *Computing Research Repository*, arXiv:2501.12948.
- DeepSeek-AI, Aojun Liu, Bingxue Feng, Bingxuan Wang, Bo Tang, Chong Chen, Chong Cheng, Junqi Dai, Zhiyong Deng, 1 others. 2024. [Deepseek-V3 technical report](#). *Computing Research Repository*, arXiv:2412.19437.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [QLoRA: Efficient fine-tuning of quantized LLMs](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115.
- Yu'na Guo, Chao Liu, Wenjing Lian, and Jie Wang. 2022. [Principles of relativity between prescription and syndrome in Treatise on Febrile Diseases](#). *Chinese Journal of Experimental Traditional Medical Formulae*, 28(22):189–195.
- Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. [Large language models are reasoning teachers](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14852–14882.
- Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. [Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *Proceedings of the International Conference on Learning Representations*.
- Miao Jiang, Cheng Lu, Chi Zhang, Jing Yang, Yong Tan, Aiping Lu, and Kelvin Chan. 2012. [Syn-drome differentiation in modern research of traditional Chinese medicine](#). *Journal of Ethnopharmacology*, 140(3):634–642.
- Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024. [RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 26874–26901. PMLR.
- Wei Long, Xiao-Dong Zhang, Hong-Ying Wu, Jin Jin, Guang-Yun Yu, Xin He, Hao Wang, Xiu Shen, Ze-Wei Zhou, Pei-Xun Liu, and Sai-Jun Fan. 2013. [Study on incompatibility of traditional Chinese medicine: Evidence from formula network, chemical space, and metabolism room](#). *Evidence-Based Complementary and Alternative Medicine*, 2013:352145.
- Meta AI, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelhammer, 1 others. 2026. [The Llama 4 herd: Architecture, training, evaluation, and deployment notes](#). *Preprint*, arXiv:2601.11659. Withdrawn from arXiv due to incorrect authorship.
- OpenAI. 2025. Introducing GPT-5. <https://openai.com/index/introducing-gpt-5/>.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, 1 others. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Liam Castellon, Marzyeh Celi, Adji Dieng, Rishi Li, Connie Liu, Anil Kumar, Giorgia DeSalvo, Vivek Natarajan, Alan Karthikesalingam, Yossi Matias, Dale Webster,

- Greg S. Corrado, and Andrew Y. Ng. 2023. [Large language models encode clinical knowledge](#). *Nature*, 620(7972):172–180.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Connie Liu, Shashanka Eslami, Noah Poon, Mitesh Bhojwani, Chris Newman, Chris Shukla, Mahdi Mousavi, Phong Le, Yong Wu, 12 others. 2025. [Toward expert-level medical question answering with large language models](#). *Nature Medicine*, 31(3):943–950.
- Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, David Chang, Andrew Chiu, Gary DeMichele, Xiao Feng, Jon Ghosh, Aakanksha Haldar, Isabelle Hassani, Tony Kanj, Khaled Kapp, Sanmi Koyejo, S. Sara Mahdavi, Yossi Matias, Jim McLean, Subhrajit Mukherjee, 11 others. 2024. [Towards generalist biomedical AI](#). *NEJM AI*, 1(3):AIoa2300138.
- Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, 7 others. 2025. [Towards conversational diagnostic artificial intelligence](#). *Nature*, 642(8067):442–450.
- Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023. [HuaTuo: Tuning LLaMA model with Chinese medical knowledge](#). *Preprint*, arXiv:2304.06975. Model also known as BenTsao (bencao, materia medica).
- Jie Wang and Xingjiang Xiong. 2009. Connotation and principles of prescription–syndrome correspondence. *Journal of Traditional Chinese Medicine*, 50(3):197. Chinese: fangzheng duiying neihan ji yuanze tantao.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.
- Ming Xie. 2023. *Formulary Science*. Science Press, Beijing, China. Chinese: fangjixue (formulary science). National TCM higher-education textbook; discusses the unified li-fa-fang-yao framework.
- Aiyuan Yang, Bowen Xiao, Bingxuan Wang, Bolin Wang, Bin Zhou, Chang Li, Chao Hao, Dehui Huang, Dong Lyu, Fan Wei, 1 others. 2023. [Baichuan 2: Open large-scale language models](#). *Computing Research Repository*, arXiv:2309.10305.
- Yue Yu, Zixuan Jin, Fukun Luo, Wei Liu, Pengqian Wang, and Xingjiang Xiong. 2023. [Research strategy of core prescription–syndrome based on disease–syndrome–treatment integration](#). *China Journal of Chinese Materia Medica*, 48(10):2626–2633. Chinese: jiyu bing-zheng-zhi jiehe de hexin fangzheng yanjiu silu; explicitly discusses the li-fa-fang-yao clinical workflow.
- Wenjing Yue, Xiaoling Wang, Wei Zhu, Xinyi Zhang, Yukun Wang, Zhiheng Huang, Zihan Li, Yixuan Li, Teng Zhang, 1 others. 2024. [TCMBench: A comprehensive benchmark for evaluating large language models in traditional chinese medicine](#). *Preprint*, arXiv:2406.01126. ArXiv preprint.
- Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. [MedDialog: Large-scale medical dialogue datasets](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250.
- Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. [HuatuoGPT, towards taming language model to be a doctor](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10859–10885, Singapore. Association for Computational Linguistics.
- Tingting Zhang, Zonghai Huang, Yaqiang Wang, Chuanbiao Wen, Yangzhi Peng, and Ying Ye. 2022. [Information extraction from the text data on traditional chinese medicine: A review on tasks, challenges, and methods from 2010 to 2021](#). *Evidence-Based Complementary and Alternative Medicine*, 2022:1679589.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Yanzhe Qiao, Yun Zhang, Hui Zhou, Jie Chen, Junyang Chen, 1 others. 2025. [Qwen3 embedding: Advancing text embedding and reranking through foundation models](#). *Computing Research Repository*, arXiv:2506.05176.
- Zhongjing Zhang. ca. 200. *Treatise on Febrile Diseases and Miscellaneous Diseases (Shang Han Lun)*. Classic TCM text; foundational syndrome-differentiated prescribing.

A Clinical Background and Gap Motivation

This appendix provides the TCM domain knowledge underlying the three clinical gaps defined in Section 1.

A.1 The *Li-Fa-Fang-Yao* Paradigm and the SR Gap

The *li-fa-fang-yao* (理法方药) framework names the four-stage clinical reasoning chain through which TCM moves from pathomechanism to medication: *li* (pathomechanism established via syndrome differentiation), *fa* (therapeutic principle), *fang* (formula selection), and *yao* (herbal medication and compatibility) (Jiang et al., 2012). National formulary curricula treat this as the unified logic linking syndrome differentiation to prescribing (Xie, 2023), while prescription–syndrome (*fang–zheng*) correspondence research formalizes the requirement that formulas align with pathogenesis rather than symptoms alone (Wang and Xiong, 2009; Guo et al., 2022; Yu et al., 2023). Historical roots trace to Zhang Zhongjing’s *Treatise on Febrile Diseases* (*Shang Han Lun*), which codified syndrome-differentiated treatment with relatively fixed formula–syndrome relations (Zhang, ca. 200; Guo et al., 2022).

A TCM practitioner follows this structured diagnostic chain in clinical practice:

- **Syndrome differentiation** (辨证): integrates symptoms, tongue coating, and pulse qualities to identify the pathogenic pattern (e.g., Liver Qi stagnation with Spleen deficiency).
- **Treatment principle** (立法): establishes the therapeutic strategy (e.g., “course the liver and fortify the spleen”).
- **Formula construction** (选方): selects herbs under the *jun-chen-zuo-shi* (君臣佐使) hierarchy—king herb targeting the main syndrome, minister herbs assisting, assistant herbs treating secondary symptoms or reducing toxicity, and courier herbs guiding or harmonizing.
- **Safety verification** (配伍验证): checks herb–herb incompatibilities, especially the absolute prohibitions in Appendix A.3.

Current LLMs perform end-to-end symptom-to-prescription mapping, producing isolated conclusions without this reasoning chain. In clinical practice, a prescription without an auditable rationale is useless—a practitioner cannot verify its appropriateness, nor learn from or correct it. This is the **SR**

Gap. The chain must be explicit, stepwise, and follow the fixed *li-fa-fang-yao* paradigm rather than free-form text (Wang and Xiong, 2009; Yu et al., 2023).

A.2 Follow-Up Visits and the LA Gap

Follow-up visits constitute a substantial share of TCM outpatient care. In a follow-up, the patient presents with a new set of symptoms that reflect the therapeutic response to the previous prescription. The clinician must:

1. Compare the current symptoms with those recorded at the previous visit.
2. Identify which aspects have improved, which have persisted, and which new manifestations have appeared.
3. Adjust the prescription accordingly—adding herbs to address remaining or new patterns, removing herbs that are no longer needed, and modifying dosages. This is *sui zheng jia jian* (随证加减, “adding or subtracting according to the pattern”).

Errors in this step are the most common source of medication problems in TCM. Existing LLM approaches treat each visit independently (static prescribing), completely missing the longitudinal dependency. A model that cannot reason about change from the previous prescription cannot safely support follow-up care. This is the **LA Gap**.

A.3 Absolute Pharmacological Prohibitions and the SC Gap

TCM has two classic sets of herb–herb incompatibilities that are considered absolute (Long et al., 2013):

- **Shi Ba Fan** (十八反, Eighteen Incompatibilities): six core pairs with derived synonyms. The most clinically important are licorice (Gan Cao) ↔ Gan Sui, Da Ji, Yuan Hua, Hai Zao; and aconite (Chuan Wu, Cao Wu, Fu Zi) ↔ Bei Mu, Gua Lou, Ban Xia, Bai Ji.
- **Shi Jiu Wei** (十九畏, Nineteen Mutual Antagonisms): ten pairs, e.g., Ding Xiang ↔ Yu Jin; Ren Shen ↔ Wu Ling Zhi; Rou Gui ↔ Chi Shi Zhi.

These are pharmacological facts, not soft preferences. Combining such herbs can cause severe toxicity or complete loss of efficacy. LLMs that generate such pairs are clinically dangerous, yet often do so with the same confidence as safe prescriptions, and no built-in mechanism detects or blocks such outputs. However, standard RLHF treats safety

as a preference to be learned from human labels, which is costly, inconsistent, and may still miss violations. A rule-based approach directly encoding these prohibitions would be more reliable—if it can be effectively integrated into alignment. This is the **SC Gap**.

A.4 Why These Gaps Differ from General LLM Limitations

While general LLMs also suffer from hallucination and weak reasoning, the TCM context makes each gap clinically concrete:

- The **SR Gap** is not about any reasoning, but about reasoning that follows the *li-fa-fang-yao* clinical paradigm (Xie, 2023; Guo et al., 2022); free-form CoT is insufficient.
- The **LA Gap** is not about multi-turn dialogue (which LLMs handle well), but about comparing clinical states and justifying adjustments based on therapeutic feedback—causal reasoning over time.
- The **SC Gap** is not about general toxicity or harmful content, but about a small, closed set of deterministic pharmacological rules that general safety filters may miss because they are rare in general text.

These distinctions motivate the three specialized components (PG-CoT, Dynamic, K-RL) and the four-stage progressive training pipeline.

B Training Hyperparameters

Table 6 summarizes the shared training configuration across all four stages.

C Data Pipeline Details

Table 1 and Figure 5 summarize the data pipeline; dataset statistics are presented in Section 4.1.

Data source. The primary data source consists of 124,593 clinical encounter records from TCM outpatient departments, containing patient demographics, chief complaints, case histories, tongue and pulse descriptions, syndrome differentiations, and herbal prescriptions. Auxiliary resources include a TCM pharmacological knowledge base with standard herb names, aliases, properties (nature, flavor, channel entry), and toxicity classifications, along with a classical formula analysis compendium.

Entity alignment. Clinical prescriptions use diverse herb names and preparation forms (e.g., 黑顺片 for Fu Zi, 炒白术 for processed Bai Zhu). We

normalize herb names through a three-tier matching strategy: exact match against standard names and aliases; prefix stripping of 18 common preparation forms (e.g., 炒/chao, 炙/zhi, 制/zhi, 煨/duan) followed by re-matching; and manual review of unmapped entries. This achieves 95.76% coverage of clinical herb names, with 10,106 records lacking valid prescriptions excluded.

CoT distillation pipeline. We design the structured reasoning schema following the *li-fa-fang-yao* paradigm and employ an LLM-auxiliary, expert-verified pipeline to construct training data. Specifically, we use DeepSeek-R1 (DeepSeek-AI et al., 2025) via the SiliconCloud Batch API (temperature $\tau=0.6$, max_tokens=4096) to generate reasoning drafts in JSON format for each case record. These drafts then undergo extensive human review and correction: TCM practitioners verify syndrome differentiation accuracy, validate treatment principle–formula alignment, and correct pharmacological reasoning errors. This human-in-the-loop process ensures that the training data reflects clinically sound reasoning rather than unconstrained model generation. The pipeline yields 42,090 verified R1 reasoning samples across 9 batches (17,401 static + 24,689 dynamic) prior to train/test decontamination. The student model is then fine-tuned on these human-verified samples via supervised learning, learning to produce the structured reasoning chain before the prescription. This transforms the training objective from $P(y | x)$ to $P(r, y | x)$, providing stronger gradient signal through intermediate supervision. Final post-decontamination training counts (16,530 PG-CoT and 39,158 Dynamic samples) are reported in Table 1.

Longitudinal extraction. Because the clinical dataset lacks unique patient identifiers, we construct pseudo-IDs using a two-tier strategy:

- Records with detailed case history (> 20 characters): ID = gender + hash(history[: 100])
- Records with brief history: ID = gender + inferred_birth_year

This yields 17,985 patient trajectories containing 109,615 total visits from 124,593 raw records. Single-visit patients (14,978) are excluded as they carry no follow-up signal.

We classify prescription changes between con-

Item	Setting
GPU	1× NVIDIA A800-SXM4-80GB
Quantization	QLoRA 4-bit (NF4) (Dettmers et al., 2023)
Framework	Unsloth
LoRA targets	q, k, v, o -proj, gate, up, down-proj
LoRA α / dropout	16 / 0
<i>SFT (Stages 1–3)</i>	
Epochs	3
Optimizer	AdamW 8-bit
LR scheduler	Linear
Max length	2,048
<i>DPO (Stage 4)</i>	
Epochs	1
DPO β	0.1
Learning rate	5×10^{-6}
Max prompt length	1,024
<i>Scale-adaptive</i>	
LoRA r : <7B / 7–14B / >14B	16 / 32 / 64
LR: <7B / 7–14B / >14B	2×10^{-4} / 1×10^{-4} / 5×10^{-5}
Batch/GPU: <7B / 7–14B / >14B	4 / 4 / 2
Grad accum: <7B / 7–14B / >14B	4 / 4 / 8
<i>Additional details</i>	
Weight decay	0.01 (SFT only)
Warmup	5 steps (SFT) / 10% (DPO)
Precision	bf16 / fp16
Gradient checkpointing	Unsloth Smart
Random seed	3407

Table 6: Training configuration. Larger models use higher LoRA rank and lower learning rates to prevent catastrophic forgetting, with unified effective batch size of 16. DPO is initialized from the Dynamic SFT adapter to preserve reasoning and longitudinal capabilities.

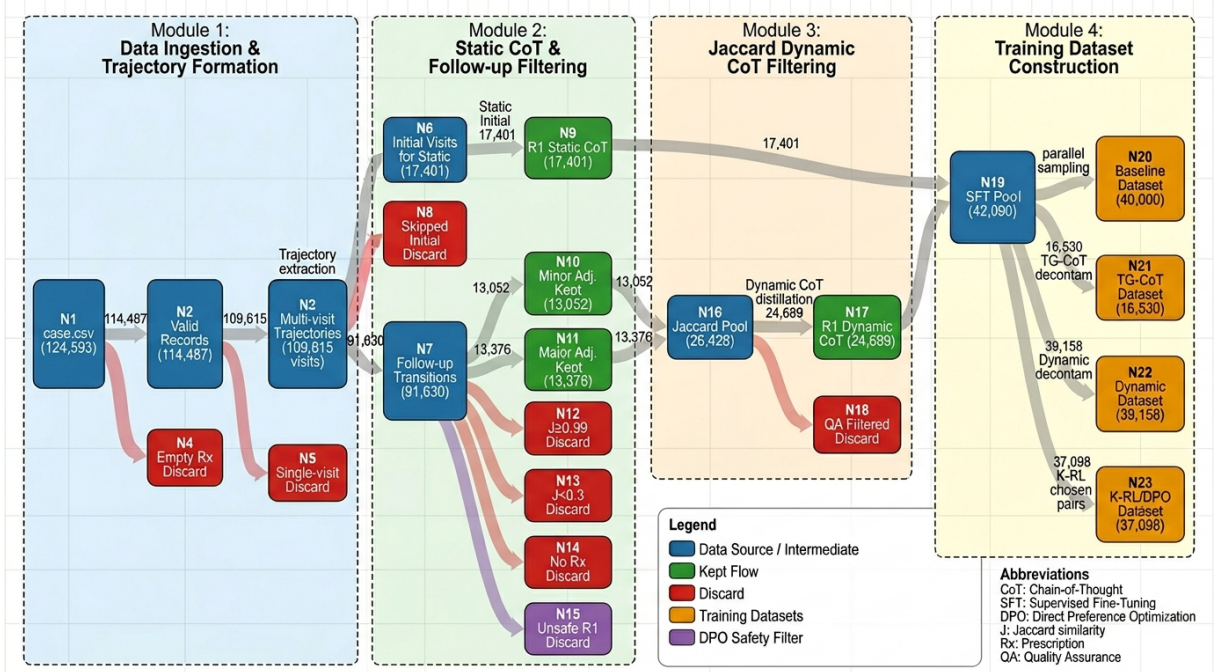


Figure 5: Data pipeline from 124,593 raw clinical records to the four training datasets. Trajectory extraction yields 17,985 patients (109,615 visits = 17,401 valid initial visits + 91,630 follow-up transitions + 584 skipped initial visits). Jaccard-based filtering retains 26,428 follow-up transitions; after R1 distillation and expert QA, 24,689 dynamic CoT samples join 17,401 static samples in a 42,090-sample SFT pool. Red streams denote discards; the four training datasets are parallel splits of this pool.

secutive visits using Jaccard similarity of herb sets:

$$J(\mathbf{y}^{(t)}, \mathbf{y}^{(t+1)}) = \frac{|\mathcal{H}^{(t)} \cap \mathcal{H}^{(t+1)}|}{|\mathcal{H}^{(t)} \cup \mathcal{H}^{(t+1)}|} \quad (7)$$

where $\mathcal{H}^{(t)}$ denotes the herb set at visit t . We retain follow-ups with $0.3 \leq J < 0.9$:

- $J \geq 0.99$ (continuation): prescription nearly unchanged, carrying no adjustment signal.
- $0.7 \leq J < 0.9$ (minor adjustment): small additions/removals, the most common clinical scenario.
- $0.3 \leq J < 0.7$ (major adjustment): substantial formula change, indicating disease progression or therapeutic pivot.
- $J < 0.3$ (complete change): likely unrelated prescription, too noisy for training.

This filtering yields 24,689 dynamic follow-up samples from 91,630 total follow-up transitions.

D PG-CoT Pilot Experiment: Free-form vs. Paradigm-Guided CoT

Section 3.2 motivates PG-CoT by contrasting paradigm-guided reasoning against free-form CoT. Here we report the pilot experiment that quantifies this contrast.

Setup. We use Mistral-7B as the student model and the same DeepSeek-R1 teacher to distill two training sets of equal size from the same 500 held-out clinical cases. Training set A (**Free-form CoT**) contains R1-generated reasoning drafts without structural constraints: the model is free to produce any narrative chain before the prescription. Training set B (**PG-CoT**, ours) contains R1 drafts that are subsequently structured and verified against the *li-fa-fang-yao* schema (syndrome analysis, treatment principle, formula rationale, safety verification), following the same expert-verified pipeline described in Section 3.2. Both sets are then fine-tuned on Mistral-7B under identical hyperparameters (Table 6), and evaluated on the same unified test set of 871 cases.

Metrics. We compare the two variants on three dimensions: (1) **PQS** for prescription quality; (2) **Clinician-rated auditability**, the proportion of reasoning steps that a TCM practitioner can verify against the clinical record (rated by two licensed TCM practitioners on a random subset of 100 outputs); and (3) **Logical inconsistency rate**, the proportion of outputs containing at least one reasoning step that contradicts a prior step (e.g., a

cold-syndrome diagnosis followed by heat-clearing herbs).

Results. Table 7 presents the results. PG-CoT achieves substantially higher auditability (+0.29) and lower logical inconsistency (−0.12) than free-form CoT, while also improving prescription quality. This confirms that the *li-fa-fang-yao* paradigm constraint serves as a performance-enhancing inductive bias rather than a limitation on model expressiveness.

Variant	PQS ↑	Auditability ↑	Inconsistency ↓
Free-form CoT	0.423	0.33	0.23
PG-CoT (ours)	0.651	0.62	0.11

Table 7: Pilot experiment: Free-form CoT vs. PG-CoT on Mistral-7B. Auditability and Inconsistency are assessed via human comparison by two licensed TCM practitioners on a random subset of 100 outputs. Auditability = proportion of verifiable reasoning steps; Inconsistency = proportion of outputs with at least one self-contradictory step. Both variants use the same training data source and hyperparameters.

E Complete Model List

Table 8 lists all models tracked by the automated pipeline.

F Metric Design Rationale

PQS component selection and weighting. PQS is co-designed with licensed TCM practitioners to reflect clinical priorities in prescription evaluation. Dosage accuracy (Acc_{dose} , 35%) receives the highest weight because TCM adheres to the principle that “dosage determines efficacy”: the same herb at different dosages can produce opposing therapeutic effects (e.g., Da Huang at low dose stops diarrhea, at high dose promotes purgation). Herb-set F1 ($F1_H$, 30%) captures whether the correct herbs are selected, the primary clinical concern after dosage. Herb-set Jaccard (J , 15%) supplements F1 with set-level overlap, penalizing both omission and commission. Herb count reasonableness (OK_{cnt} , 10%) flags structurally abnormal prescriptions (e.g., single-herb or 30-herb outputs), as TCM formulas typically contain 5–15 herbs. The safety score (S , 10%) provides a hard constraint gate: prescriptions with contraindication violations receive $S < 1$, ensuring that safety violations directly reduce PQS even when other components are high. The relatively lower weight of S reflects its role

Model Key	HuggingFace ID	Params	Family
qwen3_5_0.8b	unsloth/Qwen3.5-0.8B	0.8B	Qwen
qwen3_5_2b	unsloth/Qwen3.5-2B	2B	Qwen
qwen3_5_9b	unsloth/Qwen3.5-9B	9B	Qwen
qwen3_5_27b	unsloth/Qwen3.5-27B	27B	Qwen
gemma2_2b	unsloth/gemma-2-2b-it	2B	Gemma
gemma2_9b	unsloth/gemma-2-9b-it	9B	Gemma
gemma2_27b	unsloth/gemma-2-27b-it	27B	Gemma
phi3_mini	unsloth/Phi-3-mini-4k-instruct	3.8B	Phi
llama3_1_8b	unsloth/Meta-Llama-3.1-8B-Instruct	8B	LLaMA
mistral_7b	unsloth/Mistral-7B-Instruct-v0.3	7B	Mistral
mistral_24b	unsloth/Mistral-Small-24B-Instruct-2501	24B	Mistral
deepseek_r1_8b	unsloth/DeepSeek-R1-Distill-Llama-8B	8B	LLaMA
baichuan2_7b	baichuan-inc/Baichuan2-7B-Chat	7B	Baichuan

Table 8: Models tracked by the automated pipeline. Baichuan2-7B is compared as a TCM-specific zero-shot baseline (Table 4) rather than a fine-tuned model.

as a necessary precondition rather than a discriminative quality signal: on the standard test set, S saturates to 1.0 across stages, and its discriminative value lies in the adversarial VR metric rather than in PQS itself.

Why CQS introduces A_{reason} . PQS measures whether the prescription is correct, but cannot assess whether the model arrived at that prescription through clinically valid reasoning. This distinction is central to our SR Gap argument: a model that produces the right prescription via spurious correlations is clinically unsafe, because such shortcuts fail on out-of-distribution cases. CQS addresses this by introducing the reasoning alignment score A_{reason} , which quantifies how closely the model’s structured *li-fa-fang-yao* reasoning chain matches the expert-verified reference chain. The 0.6/0.4 weighting between PQS_{base} and A_{reason} reflects the primacy of prescription quality as the clinically verifiable outcome: a correct prescription that cures the patient is the ultimate clinical goal, and a correct prescription with auditable reasoning further enables practitioner oversight. We use the format-stable baseline SFT PQS (PQS_{base}) rather than stage-specific PQS across all stages to avoid conflating format bias artifacts (Appendix H) with reasoning quality changes.

A_{reason} implementation.

$$\begin{aligned} \text{CQS} &= 0.6 \cdot \text{PQS}_{\text{base}} + 0.4 \cdot A_{\text{reason}}, \\ A_{\text{reason}} &= \sum_k w_k \cdot \cos(\mathbf{e}_k^{\text{gt}}, \mathbf{e}_k^{\text{pred}}) \end{aligned} \quad (8)$$

Section weights w_k : syndrome analysis / evolution (30%), treatment principle / adjustment (20%), formula rationale (35%), safety check (15%).

A_{reason} is computed using Qwen3-Embedding-8B (Zhang et al., 2025), loaded locally via sentence-transformers. For each CoT output, the four structured sections are extracted by JSON parsing and encoded; cosine similarity to the ground-truth section embeddings is weighted (0.30, 0.20, 0.35, 0.15) and summed to yield A_{reason} . CQS is then $0.6 \text{PQS}_{\text{base}} + 0.4 A_{\text{reason}}$ on the same test instance. Outputs that fail JSON parsing receive $A_{\text{reason}} \approx 0$ on missing sections, which lowers CQS accordingly.

CQS interpretation caveat. CQS uses the format-stable baseline SFT PQS (PQS_{base}) as a fixed constant across all stages; cross-stage CQS improvements therefore reflect *only* changes in A_{reason} , not in prescription quality. This design avoids conflating format bias artifacts with reasoning quality (Appendix H), but it also means CQS is insensitive to genuine prescription quality degradation. For example, Mistral-7B PQS declines from 0.701 (baseline SFT) to 0.350 (K-RL) due to format bias, yet CQS rises from 0.728 to 0.766—the improvement is entirely attributable to the A_{reason} term. Readers should consult stage-specific PQS values (Table 9) alongside CQS for a complete picture of model behavior. We further note that A_{reason} measures semantic similarity to the R1-distilled reference reasoning, not clinical reasoning correctness per se; a model that independently discovers an equally valid but different diagnostic pathway would be penalized. The expert evaluation (Appendix K) partially addresses this by calibrating automatic scores against human judgment.

Why VR is proposed. The rule-based safety score S embedded in PQS serves as a necessary

precondition for prescription validity, but on the standard held-out test set it saturates to 1.0 across all stages, providing no stage- or architecture-level discrimination on safety. This saturation occurs because standard test cases do not actively challenge model safety boundaries. VR is designed to probe the model’s ability to *maintain safety under adversarial pressure*: by constructing targeted prompts that explicitly or implicitly elicit contraindication violations, VR measures how robustly a model upholds pharmacological constraints when pressured to violate them. This adversarial framing provides the discriminative safety signal that *S* cannot: VR ranges from 0.038 to 0.432 across our evaluation (Table 3), revealing substantial stage- and architecture-dependent variation that would be invisible under the standard test paradigm. The three probe categories (direct requests, contextual triggers, open generation) are designed to test safety maintenance across increasing levels of subtlety, from explicit rule violation requests to scenarios where contraindications must be recognized without prompting.

G Complete Results

Table 9 presents the complete evaluation across all 12 fine-tuned models and four training stages.

DS-R1-Distill-8B anomaly. DS-R1-Distill-8B exhibits an unusual PQS trajectory ($0.560 \rightarrow 0.268 \rightarrow 0.649 \rightarrow 0.605$): the PG-CoT stage produces a sharp PQS drop followed by a Dynamic-stage surge above Baseline. This pattern is explained by its extremely low Structural Completeness (StrC = 0.09): the model’s PG-CoT outputs fail to conform to the JSON schema, causing parser mis-extraction and severely deflating PQS (an extreme form of the format bias discussed in Appendix H). Dynamic SFT partially resolves this formatting issue, enabling proper prescription extraction and restoring PQS above Baseline. The CQS trajectory ($0.566 \rightarrow 0.570 \rightarrow 0.597$) is far more stable, confirming that the PQS volatility is a parser artifact rather than a genuine quality shift.

H Format Bias Quantification

Table 10 quantifies the format bias on PG-CoT outputs: dosage-pattern-only extraction from the *fang-yi-jie-gou* section raises Herb F1 from ~ 0.26 to ~ 0.73 and reduces zero-F1 samples from $\sim 400/871$ to $\sim 10/871$, confirming that the appar-

ent PQS decline associated with CoT is a parser artifact rather than a genuine quality drop.

The magnitude of format bias varies substantially across model families. All four models in Table 10 exhibit similar standard Herb F1 (~ 0.25 – 0.28) and Rx-Only recovery (~ 0.73 – 0.76), suggesting that the extraction error rate is relatively uniform among models that produce well-formed structured outputs. However, models with low Structural Completeness (StrC, Table 9) experience more severe bias: DS-R1-Distill-8B (StrC=0.09) shows a PQS swing of $0.560 \rightarrow 0.268$ at the PG-CoT stage, while Qwen3.5-27B (StrC=0.06) drops from $0.566 \rightarrow 0.310$. The CQS metric mitigates this by using the format-stable baseline PQS, ensuring that reasoning quality rather than parsing artifacts drives cross-stage comparisons.

I Small Model Structured Output Failures

Some models fail to produce valid structured output parseable into the four required *li-fa-fang-yao* sections (syndrome, principle, formula, safety). The failure mode is consistent: affected models begin the JSON structure correctly but produce verbose, repetitive text in the *fang-yi-jie-gou* section (e.g., Gemma-2-2B repeating “性味凉” (*xing wei liang*, “nature and flavor cool”) dozens of times), exceeding the generation length budget and leaving the output without a closing brace. This makes the output unparseable as JSON rather than failing to learn the reasoning structure—the syndrome and treatment principle sections are typically well-formed. Interestingly, scale alone does not predict success: Qwen3.5-0.8B and -2B achieve high A_{reason} (>0.93) and CQS ≈ 0.71 , DS-R1-Distill-8B reaches CQS ≈ 0.57 – 0.60 across CoT stages, while Gemma-2-27B remains low, indicating that format compliance depends on model architecture rather than parameter count. Qwen3.5-27B also exhibits low Structural Completeness (StrC = 0.06), suggesting that larger models in families with pre-existing format issues do not automatically resolve them.

J Case Studies

We present two representative cases from the test set, showing outputs from Mistral-24B across all four training stages.

Case 1: Initial visit with spleen qi deficiency and dampness. The patient presents with abdominal

Model	Scale	PQS				Herb F1				CQS			StrC	Safety
		Base	CoT	Dyn	KRL	Base	CoT	Dyn	KRL	CoT	Dyn	KRL	Dyn	All
<i>Small models (< 7B)</i>														
Qwen3.5-0.8B	0.8B	0.639	0.351	0.363	0.329	0.476	0.236	0.248	0.216	0.707	0.723	0.716	0.99	1.0
Qwen3.5-2B	2B	0.699	0.390	0.400	0.363	0.531	0.268	0.277	0.291	0.711	0.715	0.702	1.0	1.0
Gemma-2-2B	2B	0.604	0.395	0.405	0.420	0.454	0.266	0.264	0.294	0.387	0.376	0.553	0.32	1.0
Phi-3-Mini	3.8B	0.643	0.401	0.367	0.380	0.406	0.306	0.291	0.314	0.422	0.442	0.501	0.16	1.0
<i>Medium models (7–14B)</i>														
Qwen3.5-9B	9B	0.449	0.321	0.294	0.213	0.195	0.186	0.195	0.145	0.584	0.522	0.519	0.00	1.0
LLaMA-3.1-8B	8B	0.710	0.400	0.385	0.390	0.530	0.278	0.256	0.266	0.727	0.731	0.724	1.0	1.0
Mistral-7B	7B	0.701	0.405	0.391	0.350	0.519	0.276	0.263	0.225	0.728	0.732	0.766	1.0	1.0
DS-R1-Distill-8B	8B	0.560	0.268	0.649	0.605	0.290	0.189	0.583	0.515	0.566	0.570	0.597	0.09	1.0
Gemma-2-9B	9B	0.717	0.377	0.382	0.365	0.547	0.253	0.260	0.240	0.721	0.728	0.729	1.0	1.0
<i>Large models (> 14B)</i>														
Qwen3.5-27B	27B	0.566	0.310	0.375	0.361	0.503	0.322	0.319	0.388	0.572	0.577	0.581	0.06	1.0
Gemma-2-27B	27B	0.683	0.388	0.391	0.379	0.510	0.260	0.258	0.248	0.676	0.683	0.701	0.07	1.0
Mistral-24B	24B	0.721	0.380	0.375	0.393	0.550	0.257	0.255	0.277	0.723	0.718	0.731	1.0	1.0

Table 9: Complete results across all evaluated models. StrC = Structural Completeness (for Dynamic stage); Safety score is 1.0 across all models and experiments.

Model	Herb F1		Rx-Only	0-F1 Rec.
	Std	Rx-Only	PQS	(#)
Mistral-24B	0.257	0.727	0.513	394
LLaMA-3.1-8B	0.278	0.746	0.523	350
Mistral-7B	0.276	0.758	0.528	357
Gemma-2-9B	0.253	0.740	0.520	385

Table 10: Format bias quantification on PG-CoT outputs. “Std” = standard pipeline; “Rx-Only” = dosage-pattern-only extraction; “0-F1 Rec.” = number of zero-Herb-F1 samples recovered.

distension, loose stools, fatigue, pale tongue with white coating, and a weak pulse.

Baseline: The model outputs a prescription directly:

党参15g 白术12g 茯苓15g 甘草6g 陈皮9g 半夏9g 砂仁6g
Dang Shen 15g, Bai Zhu 12g, Fu Ling 15g, Gan Cao 6g, Chen Pi 9g, Ban Xia 9g, Sha Ren 6g

No diagnostic rationale is provided; the practitioner cannot verify why these herbs were selected.

PG-CoT: The model produces a structured reasoning chain followed by the prescription:

[Syndrome] 脾气虚兼湿—腹胀便溏，乏力，舌淡苔白，脉弱，符合脾气虚弱、运化失司之证。
Spleen qi deficiency with dampness—abdominal distension, loose stools, fatigue, pale tongue with white coating, weak pulse, consistent with spleen qi weakness and impaired transportation.
 [Principle] 健脾益气，化湿和中。
Fortify the spleen and replenish qi, transform dampness and harmonize the middle.
 [Formula] 以四君子汤健脾益气为君，二陈

汤化湿为臣，砂仁醒脾为佐。
Si Jun Zi Tang as king to fortify spleen and replenish qi; Er Chen Tang as minister to transform dampness; Sha Ren as assistant to revive the spleen.
 [Safety] 检查十八反、十九畏：无配伍禁忌。
Checked Shi Ba Fan and Shi Jiu Wei: no contraindication combinations.
 [Prescription] 党参15g 白术12g 茯苓15g 甘草6g 陈皮9g 半夏9g 砂仁6g
Dang Shen 15g, Bai Zhu 12g, Fu Ling 15g, Gan Cao 6g, Chen Pi 9g, Ban Xia 9g, Sha Ren 6g

The reasoning is auditable: each herb’s role is justified within the *li-fa-fang-yao* paradigm.

Dynamic: Same as PG-CoT for this initial visit (no prior prescription to adjust).

K-RL: Identical reasoning chain with safety verification more prominent; prescription unchanged.

Case 2: Follow-up visit with partial improvement after initial treatment. The patient received Si Jun Zi Tang for spleen qi deficiency. Two weeks later, abdominal distension has improved but loose stools persist; new symptom: night sweating.

Baseline: The model outputs a new prescription with no reference to the prior one:

黄芪20g 白术12g 防风9g 浮小麦15g 牡蛎20g 甘草6g
Huang Qi 20g, Bai Zhu 12g, Fang Feng 9g, Fu Xiao Mai 15g, Mu Li 20g, Gan Cao 6g

Dynamic: The model produces longitudinal reasoning:

[Evolution] 腹胀减轻提示脾气渐复，但便溏未愈、新增盗汗，说明脾虚及气虚卫外不

固。

Reduced distension indicates partial spleen qi recovery, but persisting loose stools and new night sweating suggest spleen deficiency and weak defensive qi.

[Adjustment] 保留党参、白术、茯苓健脾基础；加黄芪、防风益气固表（玉屏风散意）；加浮小麦、牡蛎收敛止汗。去半夏、陈皮（湿已减，无需燥湿）。

Retain Dang Shen, Bai Zhu, Fu Ling as spleen-fortifying base; add Huang Qi and Fang Feng to replenish qi and secure the exterior (Yu Ping Feng San intent); add Fu Xiao Mai and Mu Li to astringe and stop sweating. Remove Ban Xia and Chen Pi (dampness reduced, no further drying needed).

[New Formula] 以四君子汤合玉屏风散加减，健脾固表止汗。

Modified Si Jun Zi Tang combined with Yu Ping Feng San to fortify the spleen, secure the exterior, and stop sweating.

[Safety] 检查十八反、十九畏：无配伍禁忌。

Checked Shi Ba Fan and Shi Jiu Wei: no contraindication combinations.

[Prescription] 党参15g 白术12g 茯苓15g 黄芪20g 防风9g 浮小麦15g 牡蛎20g 甘草6g
Dang Shen 15g, Bai Zhu 12g, Fu Ling 15g, Huang Qi 20g, Fang Feng 9g, Fu Xiao Mai 15g, Mu Li 20g, Gan Cao 6g

The model explicitly compares the current and prior states, justifying each herb addition and removal.

K-RL: Same dynamic reasoning with safety verification; prescription preserved as no contraindication violations are detected.

K Expert Evaluation Design

Automatic metrics measure surface-level agreement with reference outputs, but TCM prescription adequacy requires domain expert judgment. Two prescriptions may share no common herbs yet both be clinically reasonable (*tong bing yi zhi*—different treatments for the same disease), a scenario where Herb F1 and Jaccard incorrectly penalize one prescription. Similarly, CQS (via A_{reason}) measures semantic similarity to the R1-distilled reference reasoning, not the *correctness* of the reasoning itself. We therefore design a structured expert evaluation protocol to validate our automatic metrics and directly assess clinical quality.

Evaluation protocol. We randomly sample 100 cases from the unified test set (50 initial visits and 50 follow-up visits) and present outputs from all four training stages of the best-performing model (Mistral-24B by automatic metrics). Outputs are anonymized as System A/B/C/D with randomized ordering to prevent stage identification.

Two TCM professionals (licensed practitioners or graduate-level TCM researchers) independently score each output on five dimensions using a 1–5 Likert scale:

1. **Syndrome accuracy (D):** Does the syndrome differentiation match the patient’s symptoms, tongue, and pulse? (*CoT/Dynamic/K-RL only*)
2. **Li-fa-fang-yao coherence (C):** Is the reasoning chain logically connected from syndrome to treatment principle to formula to herbs? (*CoT/Dynamic/K-RL only*)
3. **Prescription adequacy (P):** Are the herb selection and dosage appropriate for the identified syndrome? (*All stages*)
4. **Safety (S):** Does the prescription contain any contraindication violations or toxicity concerns? (*All stages*)
5. **Adjustment flexibility (F):** Is the follow-up modification clinically justified, with reasonable explanation for additions/removals? (*Dynamic/K-RL only*)

Baseline outputs lack reasoning chains; dimensions D and C are marked N/A for baseline. Static cases have dimension F marked N/A. We report inter-rater agreement via Cohen’s κ , between-system comparisons via Wilcoxon signed-rank tests with Bonferroni correction, and auto-human metric correlation via Spearman’s ρ between automatic scores (PQS, CQS) and corresponding human dimensions (P, C).

Stage	D ↑	C ↑	P ↑	S ↑	F ↑
Baseline	N/A	N/A	4.7	3.9	N/A
PG-CoT	4.1	4.5	4.3	4.2	N/A
Dynamic	4.6	4.6	4.5	4.1	4.5
K-RL	4.5	4.3	4.6	4.8	4.5
Cohen’s κ	0.77				

Table 11: Expert evaluation results (1–5 Likert scale, mean across two raters and 100 cases). D = Syndrome accuracy; C = Li-fa-fang-yao coherence; P = Prescription adequacy; S = Safety; F = Adjustment flexibility. N/A = not applicable for that stage.

Inter-rater agreement is substantial (Cohen’s $\kappa = 0.77$). Several patterns emerge from Table 11. First, structured reasoning stages consistently outperform Baseline on prescription adequacy (PG-CoT 4.3, Dynamic 4.5, K-RL 4.6 vs. Baseline 4.7); the slight Baseline advantage on P reflects that Baseline outputs are concise prescription lists without the narrative overhead that can occasionally dilute herb selection focus. Second, K-RL achieves the high-

est safety score (4.8), confirming that DPO-based alignment with pharmacological rules effectively internalizes safety preferences, consistent with the VR reductions observed in Table 3. Third, Dynamic achieves the best syndrome accuracy (4.6) and coherence (4.6), suggesting that longitudinal reasoning training also sharpens diagnostic precision on initial visits. Fourth, Baseline’s safety score (3.9) is the lowest across all stages, underscoring the SC Gap: models without explicit safety mechanisms are more prone to contraindication violations. The expert evaluation thus validates the automatic metrics and confirms that each training stage contributes incrementally to clinical quality.

L Reproducibility

All experiments use a fixed random seed (3407) across all models and training stages. Training is performed on a single NVIDIA A800-SXM4-80GB GPU per job, with total compute estimated at approximately 700 A800-hours (52 training jobs averaging ~ 13 hours each). We plan to release training code, evaluation scripts, and model adapters upon publication. Clinical data cannot be publicly released due to patient privacy regulations, but we provide the TCMSafetyChecker rule set and adversarial probe set to facilitate replication of safety evaluation.

M Prompt Templates

Baseline instruction: "Please generate a TCM herbal prescription based on the patient’s symptoms."

PG-CoT / Dynamic instruction: "Please analyze the patient’s condition following the li-fa-fang-yao paradigm (syndrome differentiation, treatment principle, formula rationale, safety check) and generate an appropriate TCM herbal prescription."

Follow-up (Dynamic) instruction: "This is a follow-up visit. Based on the initial prescription and the patient’s current symptoms, analyze the condition evolution, explain the adjustment rationale, and generate a modified prescription."

N Zero-shot vs. Fine-tuned

Zero-shot evaluation reveals substantial family-wise variation rather than a uniform “near-zero” capability. Chinese-centric models (Qwen3.5 at 0.8B–9B) reach 0.34–0.51 PQS without domain fine-tuning; Qwen3.5-27B reaches 0.433, while English-centric 7–8B models (Mistral, LLaMA, DeepSeek-R1-Distill) remain near 0.11–0.20. Baseline SFT still yields large gains on many English-centric models (e.g., Mistral-7B: 0.113 $\rightarrow$ 0.701 PQS) and moderate gains on Chinese-centric ones (e.g., Qwen3.5-27B: 0.433 $\rightarrow$ 0.566), though Qwen3.5-9B is an exception where zero-shot already exceeds baseline SFT on PQS; we attribute this to the model’s strong Chinese-centric pretraining combining with a tendency to produce verbose, less structured outputs after SFT, which reduces dosage-accuracy extraction and thus PQS despite improved herb selection.

O Scaling and Model Family Analysis

Scale effects. Among baseline models, larger models generally achieve higher PQS (Mistral-24B: 0.721 > Mistral-7B: 0.701 > Qwen3.5-0.8B: 0.639). This advantage diminishes in CoT experiments, where PQS alone is dominated by output format effects. CQS mitigates this: models with successful structured output achieve ~ 0.71 – 0.73 CQS on PG-CoT despite PQS ~ 0.35 – 0.40 , because A_{reason} remains high (~ 0.94 – 0.95). Models that fail JSON schema compliance (Gemma-2-27B) show low CQS regardless of scale; others can maintain moderate-to-high CQS despite depressed stage-wise PQS.

Model family effects. Mistral and LLaMA models show the most consistent performance, with Baseline PQS ~ 0.70 and PG-CoT CQS ~ 0.72 – 0.73 . Qwen3.5-0.8B/2B match this pattern at ~ 0.71 – 0.72 CQS; Qwen3.5-27B holds ~ 0.57 CQS across CoT stages; Qwen3.5-9B holds ~ 0.52 CQS on dynamic/k_rl. Gemma-2-9B is stable (~ 0.72 – 0.73 CQS); Gemma-2-2B PG-CoT CQS (~ 0.39) remains below Baseline, while Gemma-2-27B stays low across stages.

Chinese-centric pretraining advantage? Contrary to intuition, Chinese-centric pretrained models (Qwen3.5 series) do not consistently outperform English-centric counterparts (Mistral, LLaMA) on this Chinese TCM task after domain SFT. This suggests that task-specific fine-tuning may matter

Model	PQS		Herb F1	
	Zero-shot	Base SFT	Zero-shot	Base SFT
<i>Small models (<7B)</i>				
Qwen3.5-0.8B	0.344	0.639	0.132	0.476
Qwen3.5-2B	0.472	0.699	0.171	0.531
Gemma-2-2B	0.160	0.604	0.034	0.454
Phi-3-Mini	0.108	0.643	0.004	0.406
<i>Medium models (7–14B)</i>				
Qwen3.5-9B	0.514	0.449	0.214	0.195
LLaMA-3.1-8B	0.200	0.710	0.051	0.530
Mistral-7B	0.113	0.701	0.002	0.519
DS-R1-Distill-8B	0.111	0.560	0.018	0.290
Gemma-2-9B	0.367	0.717	0.112	0.547
<i>Large models (>14B)</i>				
Mistral-24B	0.396	0.721	0.147	0.550
Gemma-2-27B	0.353	0.683	0.125	0.510
Qwen3.5-27B	0.433	0.566	0.119	0.503

Table 12: Zero-shot vs. baseline SFT on the unified test set (871 cases). Baichuan2-7B is reported in Table 4 as a TCM-specific zero-shot baseline.

more than pretraining language alone for final prescription quality.

Scale and safety. Adversarial VR remains substantial across scales (0.038–0.432 in Table 3), with no architecture achieving near-zero violation rates under all stages. Mistral-7B shows the highest baseline VR (0.432); Gemma-2-9B shows the largest PG-CoT reduction (0.340 $\rightarrow$ 0.094). Stage rankings vary by family (K-RL lowest on 3/10 models; PG-CoT on several others), so safety conclusions should be reported per architecture rather than as a single global winner.