CASETCM: A BENCHMARK FOR CASE-PRESERVING CROSS-SOURCE RETRIEVAL IN TRADITIONAL CHINESE MEDICINE

Anonymous authors
Paper under double-blind review

ABSTRACT

Question answering in traditional Chinese medicine (TCM) depends on links between disease-monograph pathology reasoning and pharmacopoeia herb grounding, and a usable answer has to remain traceable to the evidence route a clinician would recognize. Existing medical and TCM benchmarks, however, grade closed-book answers against a key: they ask whether a final string is right, not whether a system can recover the source case that underwrites it and cross between two structurally different sources while keeping that case intact. We present CASETCM, a benchmark for case-preserving cross-source retrieval in TCM. It comprises 3,646 pathology questions drawn from 17 disease monographs, graded against slot-structured gold answers taken verbatim from source fields, together with 1,237 herb-name-masked, candidate-constrained reverse-bridge probes over five books; the frozen benchmark files ship with a matching protocol of shared prompts, one parser, routing controls, a paraphrase robustness probe and an error taxonomy. Evaluating dense, sparse and graph retrieval, a structured RAG comparator and a deterministic reference system shows that the ceiling here is set by preserving the source case template rather than by retrieval scale: flat retrieval plateaus far below the reference system, typed field scoring adds nothing once candidates are bounded by the routed case, exact-key routing is paraphrase-brittle until a lexical fallback is added, and almost all remaining errors are wrong-case routing rather than answer realization. We anticipate that CASETCM will support retrieval systems whose evidence routes can be inspected field by field. The benchmark files, the evaluation protocol and the reference implementation are released at `https://anonymous.4open.science/r/Case-tcm-code`.

1 INTRODUCTION

Large language models (LLMs) create new opportunities to expand access to specialized medical knowledge, including culturally specific systems (Singhal et al., 2023; 2025; Liu et al., 2025b; Wang et al., 2024b; Fast et al., 2024), but their use in medicine depends on more than fluent generation: recommendations must remain traceable to the evidence route that clinicians recognize as valid. This requirement is especially strong in traditional Chinese medicine (TCM), where diagnosis and treatment depend on syndrome differentiation, pathogenesis interpretation, treatment-principle selection, formula choice, and herb-level modification rather than on disease labels alone (Chung et al., 2016; Pan et al., 2024; Dai et al., 2024; Ren et al., 2025; Wang et al., 2025). The requirement is also structural. Disease monographs preserve case-like pathology templates that link symptoms, syndrome, pathogenesis, therapy, formula, and conditional modifications, whereas the Chinese Pharmacopoeia organizes herb knowledge around efficacy, indications, dosage, and cautions (Xu et al., 2021;?). Clinicians move between the two sources, reasoning about a syndrome-pattern question in a monograph and then consulting pharmacopoeia evidence to justify a herb-level recommendation, so TCM question answering is a cross-source evidence-routing problem rather than only a retrieval problem.

Retrieval-augmented generation (RAG) supplies a general mechanism for grounding LLMs in external knowledge (Lewis et al., 2020; Gao et al., 2023), and structured retrieval improves reliability and reduces hallucination (Xiong et al., 2024a; Liu et al., 2025a; Edge et al., 2024; Gutiérrez et al., 2024; Feng et al., 2026; Hu et al., 2026). Such systems, however, organize evidence around topical

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

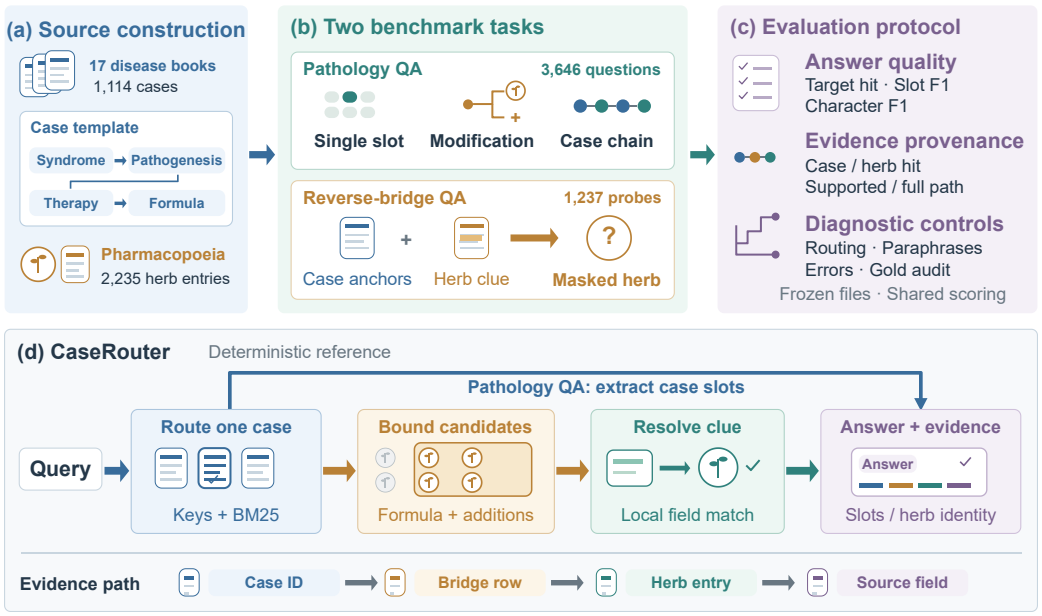

Figure 1: Overview of CASETCM. (a) Construction: 17 disease monographs are normalized into 1,114 typed cases that keep the monograph case template, and pharmacopoeia herb entries are linked through explicit bridge rows. (b) Two benchmark targets: within-book pathology QA (3,646 questions over single-slot, modification-condition, and chain families) and herb-name-masked reverse-bridge QA (1,237 probes). (c) Evaluation protocol: answer quality, evidence provenance, and diagnostic controls, on frozen files with a shared scoring parser. (d) The deterministic reference system (CASEROUTER) routes a query to one case by key lookup with a lexical BM25 fallback, bounds candidates by formula and matched additions, and returns the answer with its full evidence path (case ID → bridge row → herb entry → source field).

relevance, graph neighborhoods, or generic entity relations, which is insufficient here: a system may retrieve the right symptom vocabulary while losing the therapy principle, or the right herb description while detaching it from the pathology case that justifies its use. The dominant failure mode is evidence drift across a source-native reasoning chain.

What existing benchmarks do not measure. Chinese medical benchmarks and licensed-physician question banks (Wang et al., 2024c; Liu et al., 2023) are closed-book exams: they grade a final choice or answer against a key, so a system is never required to recover a source document's case template, and a correct string reached through the wrong evidence route costs nothing. General retrieval and knowledge-intensive benchmarks (Thakur et al., 2021; Petroni et al., 2021) do score provenance, but over one shared corpus and at page or passage granularity, so no source-native case scaffold has to survive retrieval. To our knowledge no existing resource isolates *case-preserving cross-source retrieval*: recovering the source field that underwrites an answer, from the correct case, across two structurally different sources, with the evidence route itself part of the score. We call this property retrieval-time provenance fidelity, to distinguish it from final-answer provenance.

What we provide. We introduce CASETCM, a benchmark for case-preserving cross-source retrieval in TCM, together with a deterministic reference system that instantiates the required behaviour: it routes a pathology question to a source-native case, crosses an explicit bridge into a local herb candidate set, and returns a case-grounded short answer with full-path provenance. The design principle is source-matched structure: the pathology side preserves the disease-book case template, the pharmacopoeia side is queried as an indexed herb space, and the bridge between them stays explicit and auditable (Figure 1). The reference system is deliberately deterministic, so the benchmark measures structure rather than model capacity. Our contribution is the benchmark and its evaluation protocol; the reference system is included to show that the benchmark is solvable and to expose where the difficulty lies.

We evaluate the reference system on a 17-book pathology benchmark with 3,646 question–answer pairs and a five-book reverse-bridge benchmark with 1,237 herb-name-masked, candidate-constrained questions whose formula and modification anchors remain visible. Target hit (the fraction of answers containing a gold answer alias; all endpoints are defined in Section 5.2) rises from 0.5448 for the strongest comparator to 0.8661 for the reference system, with the same ordering on both secondary endpoints and an improvement on every book. On the reverse bridge, matching the clue against the full pharmacopoeia corpus recovers under a third of the masked herbs, whereas bounding candidates by the routed case brings the reference system to near-ceiling accuracy; a raw-text control over the same candidate set matches it, so the gain comes from candidate construction rather than from typed field parsing.

Our contributions are as follows:

(i) We release CASETCM, a benchmark that separates within-book pathology QA from herb-name-masked, candidate-constrained reverse-bridge inference, with slot-structured gold answers and provenance-graded scoring.

(ii) We define an evaluation protocol for this setting, with frozen benchmark files, matched prompts, a shared parser, routing controls from random case to gold case, a paraphrase robustness probe, and an error taxonomy that separates routing failures from coverage limits.

(iii) We run existing retrieval systems under the shared protocol and show what kind of problem this is: they lose much of their accuracy by mixing evidence from several cases, and a deterministic system that keeps one case intact recovers most of that loss, so the bottleneck is how evidence is organized rather than how much is retrieved or how well it is generated.

2 RELATED WORK

2.1 RETRIEVAL-AUGMENTED AND STRUCTURED GENERATION

RAG grounds parametric generation in external knowledge (Lewis et al., 2020; Gao et al., 2023), and medical RAG has been adapted to exam-style QA, iterative evidence seeking, guideline interpretation, hepatology support, diabetes education, EHR summarization, and medical fitness assessment (Xiong et al., 2024a;b; Jeong et al., 2024; Zakka et al., 2024; Kresevic et al., 2024; Wang et al., 2024a; Alkhalaf et al., 2024; Ke et al., 2025). A systematic review maps these reported gains to retrieval architecture and strategy (Liu et al., 2025a). On the structure side, GraphRAG organizes evidence through graph communities (Edge et al., 2024); HippoRAG emphasizes graph-structured long-term memory (Gutiérrez et al., 2024); Hyper-RAG shows that hypergraphs capture higher-order medical correlations (Feng et al., 2026); and Cog-RAG separates entity-level from theme-level organization with a dual-hypergraph design (Hu et al., 2026). Hypergraph learning itself has a longer history, from regularization (Zhou et al., 2006) to hypergraph neural networks and their variants (Feng et al., 2019; Gao et al., 2020). Such systems establish that structural organization improves evidence quality, but they stay inside one literature space or a generic graph neighborhood, and none is evaluated on whether a source-native case scaffold survives retrieval. CASETCM targets that question, with a reference system that preserves two source-native structures and makes the cross-source bridge an explicit, routed object.

2.2 BENCHMARKS FOR RETRIEVAL AND MEDICAL QUESTION ANSWERING

Retrieval benchmarks standardize comparison across heterogeneous corpora: BEIR evaluates zero-shot transfer over 18 datasets and shows that in-domain strength does not imply cross-corpus strength (Thakur et al., 2021), while KILT grounds every task in one Wikipedia snapshot and requires a provenance page for each prediction (Petroni et al., 2021). Both make provenance a first-class target, which we adopt, but both assume a single shared corpus, so no source-native case scaffold has to survive retrieval and no cross-source route is scored. Medical evaluation has likewise moved past exam accuracy toward grounding and clinical validity (Xiong et al., 2024a; Ru et al., 2024), adding physician comparison, multimodal stress testing, tailored care, and workflow validation (Takita et al., 2025; Kim et al., 2025), while TCM-specific work motivates reasoning-aware digital support (Pan et al., 2024; Dai et al., 2024; Ren et al., 2025; Liu et al., 2025b; Wang et al., 2025). Chinese medical benchmarks such as CMB (Wang et al., 2024c) and CMExam (Liu

Table 1: Positioning against existing Chinese medical benchmarks. "Case template": the gold answer must preserve a monograph's syndrome–pathogenesis–therapy–formula scaffold. "Provenance": the evidence route is scored, not only the final string. CASETCM draws on disease monographs and on the Chinese Pharmacopoeia.

Benchmark	Domain	Item type	Case template	Provenance
CMB	clinical	MCQ + consultations	no	no
CMExam	licensing exam	MCQ + explanations	no	no
CASETCM	TCM monographs	slot-structured QA	yes	yes

et al., 2023), together with licensed-physician question banks, target exam-style factual recall—CMB covers multiple-choice questions and consultation cases, CMExam provides 60K+ licensing-exam questions—yet both grade the final answer only. CASETCM instead requires preserving a monograph's case scaffold, grades the evidence route as well, and reports an error taxonomy that separates routing failures from benchmark-coverage limits (Table 1).

3 PROBLEM FORMULATION

Sources, task, and gold. Disease monographs preserve case-like pathology templates that link symptoms, syndrome, pathogenesis, therapy, formula, and conditional modifications, whereas the Chinese Pharmacopoeia (Xu et al., 2021) organizes herb knowledge around efficacy, indications, dosage, and cautions. We write $\mathcal{G}_{\mathrm{path}}$ for the pathology side and $\mathcal{G}_{\mathrm{pharm}}$ for the pharmacopoeia side. The retrievable realization of the pathology side is a typed case memory $\mathcal{M}_{\mathrm{case}} = \{c_1, \ldots, c_n\}$ of normalized case fields, the pharmacopoeia side is an indexed herb space, and an explicit bridge set $\mathcal{B}_{\mathrm{bridge}}$ links a pathology-side trigger to an herb-side payload; this notation fixes the schema shared by every compared system.

Given a question q over one book, a system returns a short answer y with evidence context C. The gold set $G = \{a_1, \ldots, a_K\}$ is not a free-form reference string but a set of aliases taken verbatim from a source field: the syndrome field for a syndrome question, a condition–herb pair for a modification question, or an ordered syndrome–pathogenesis–therapy–formula scaffold for a chain question. A case-preserving system must therefore (i) route q to a source-native case $c^\star$, (ii) when the question crosses sources, bound the herb candidates through $\mathcal{B}_{\mathrm{bridge}}$ before selecting among them, and (iii) return the source field together with the evidence path that produced it.

Scoring contract. Three properties follow. Answers are graded against slot-structured gold, so a fluent response that mixes adjacent cases loses credit; the two settings below are scored separately, so case preservation is not confounded with cross-source grounding; and provenance is graded alongside the answer, so a correct string taken from the wrong source does not score. We call this retrieval-time provenance fidelity: the evidence route is part of the target, not an auxiliary diagnostic. In pathology-side disease-book QA, correctness requires recovering the case that carries the requested field. In pathology-to-pharmacopoeia reverse-bridge QA, the herb name is masked while the formula and modification anchors remain visible, so the probe evaluates candidate-bounded literature recovery rather than open-world herb identification.

Relation to passage-level provenance. Requiring provenance is not new: knowledge-intensive benchmarks already ask a system to return the page or passage that supports an answer. The requirement here is stronger in two ways, and both follow from what a TCM answer has to be. First, the supporting unit is a typed field of one specific source case rather than an arbitrary span, so a system cannot satisfy it by returning a passage that happens to contain the answer string. Second, the route has to survive the cross-source step: the answer must carry the case it came from, the bridge row it crossed, and the pharmacopoeia field it used. A correct string reached through the wrong case therefore fails even when some retrieved passage contains it, which is exactly the failure that answer-only grading cannot see.

Table 2: Composition of CASETCM. The pathology benchmark and the reverse-bridge benchmark are separate targets; the pathology question count excludes the reverse-bridge probes, and the reverse-bridge books reuse their own case files.

Pathology benchmark		Question families	
Disease books	17	Single-slot	3,060
OCR source chunks	2,699	Modification-from-condition	339
Normalized cases	1,114	Symptom-to-formula chain	247
Distinct syndrome values	543	Questions per book	62–458
Distinct pathogenesis values	222	Reverse-bridge books	5
Distinct formula names	632	Reverse-bridge probes	1,237
Questions	3,646	formula-clue / condition-clue	1,033 / 204
Pharmacopoeia herb entries	2,235	Bridge rows / distinct herbs	1,751 / 545

4 THE CASETCM BENCHMARK

4.1 MULTI-SOURCE CORPUS AND SCHEMA

Disease-book OCR output was normalized into case-centered text units: for each book we extracted structured records for disease, syndrome, pathogenesis, symptom, sign, tongue and pulse cues, therapy, treatment principle, formula and its composition, herbs, dosage, cautions, and modification conditions (full schema in the supplementary material), and questions were derived from these normalized slots so that evaluation probes literature-meaningful retrieval units rather than arbitrary spans. Herb monographs from the Chinese Pharmacopoeia (Xu et al., 2021) were parsed into the same field-structured form, and each bridge row stores a pathology-side trigger together with its herb-side payload (efficacy, indications, dosage, cautions). Construction is deterministic and reproducible from the released scripts (see the reproducibility statement), and all benchmark files were frozen before any comparative run.

4.2 INSTANCE GENERATION

Pathology instances are generated from the typed slots: single-slot questions request one field of a case, modification-from-condition questions ask what a condition adds to a base formula, and symptom-to-formula chain questions require the ordered syndrome–pathogenesis–therapy–formula scaffold. Every question comes from a slot of one specific case, and that case rather than a topic is the routing target, so a topically plausible answer drawn from a neighbouring case is scored wrong by construction. For bridge instances we mask the herb name in a pathology-to-pharmacopoeia question and expose only a short pharmacopoeia clue, with the herb identity as gold. Because the query and corpus still expose the formula and condition anchors, candidate herbs are bounded by the routed case while the clue is often globally ambiguous across the pharmacopoeia corpus, which is why we call this candidate-constrained reverse-bridge inference rather than open-world herb generation.

4.3 VALIDATION AND QUALITY CONTROL

Automatically, 99.3% of items have a gold alias that occurs in that book's case text, and the remaining 0.7% are reported as a coverage limit in Section 6.4. By hand, two annotators with TCM training independently audited 50 sampled items against the case fields and excerpts: gold was correct in 45 and 46 of the 47 decidable items (3 undecidable), giving 95.7% and 97.9% agreement and Cohen's $\kappa = 0.88$ on 98.0% inter-annotator agreement.

4.4 COMPOSITION AND EXAMPLE ITEMS

Table 2 gives the composition. The corpus keeps source vocabulary rather than a normalized ontology—543 distinct syndrome values, 222 pathogenesis values, 632 formula names—with OCR noise, formula aliases, and variable answer granularity preserved, and books contribute 62 to 458 questions each, so no single specialty dominates a macro average. The five bridge corpora connect $1{,}751$ pathology-side bridge rows to 545 herb names.

Task	Question	Gold answer & provenance
(a) Single-slot 风湿病	患者见：关节局部红肿热痛,兼见恶风畏冷,得温则舒,关节晨僵,活动后减轻,舌质红,苔白或黄白相间,脉弦紧或滑数。对应哪一证？请只回答证型名称。	寒湿化热证 Source: syndrome
(b) Modification 呼吸病	在《呼吸病》这本书中，风寒证若见“表寒重者”，应加哪几味药？请直接列药名。	表寒重者 → 麻黄、桂枝 Source: modification_pairs[].condition → herbs 鼻塞流涕重者 → 辛夷(包煎) 苍耳子 × Wrong condition　raw_clause excerpts
(c) Chain 肝胆病	患者见：胁肋刺痛,痛处固定而拒按,人夜更甚,面色晦暗,舌质紫暗或有瘀斑,脉弦涩。在《肝胆病》这本书中，这更偏什么证？其病机、治法、方药分别是什么？请按“证型：...；病机：...；治法：...；方药：...”作答。	证型：血瘀阻络证； 病机：肝郁日久,气滞血瘀,或阴伤血滞,脉络瘀阻； 治法：活血化瘀,通络止痛； 方药：血府逐瘀汤加减 Source: syndrome → pathogenesis → therapy → formula_name

Figure 2: Three verbatim CASETCM items, one per answer family, with the provenance of each gold answer; questions are shown exactly as released, including residual OCR noise. The right column records which typed field or structured record the gold was taken from: the `syndrome` field in (a), a condition–herb pair in (b), and the ordered syndrome, pathogenesis, therapy, and formula fields in (c). The grey row in (b) is the competing condition carried by the same raw clause, so answering its herbs is topically plausible but scored wrong.

Figure 2 shows what the scoring rewards; a topically plausible answer drawn from a neighbouring case scores as wrong.

5 EVALUATION PROTOCOL

5.1 TASK TAXONOMY

The pathology benchmark separates three families. Single-slot items ($N = 3{,}060$) request one field of a routed case and mainly test whether the correct case is recovered. Modification-from-condition items ($N = 339$) ask what a condition adds to a base formula, so the credited answer is the conditional delta. Symptom-to-formula chain items ($N = 247$) require the ordered syndrome–pathogenesis–therapy–formula scaffold of one case, so credit is lost as soon as evidence is mixed across cases. The reverse-bridge target adds a second axis: 1,237 herb-name-masked probes over five books (1,033 formula-clue, 204 condition-clue), with the formula and modification anchors left visible.

5.2 METRICS

We report target hit, character F1, and slot F1 as main automatic endpoints. Target hit checks whether a gold answer alias appears as a substring of the normalized answer, slot F1 is the per-slot precision and recall over slot-vocabulary mentions in the answer, and character F1 is the multiset character overlap against the best-matching gold alias. Wherever routed identifiers are available we also report gold-case hit, supported-answer rate, and full-path answer rate; supported-answer rate requires both target hit and evidence context hit, which makes retrieval-time provenance fidelity a scored endpoint rather than a side diagnostic. Averages are macro except the question-family breakdown, which is reported per question; formal definitions are in the supplementary material.

5.3 STATISTICAL PROTOCOL

Book-level comparisons use two-sided Wilcoxon signed-rank tests across books; because the minimum possible two-sided P for 17 fully aligned paired books is 1.53×10^{-5}, we interpret the main comparison as strong but low-resolution and rely on the 17-book effect pattern rather than significance alone. Focused ablations and reverse-bridge controls use paired bootstrap 95% confidence intervals and exact McNemar tests on discordant paired outcomes.

Table 3: Macro-average pathology QA comparison across 17 disease books. All rows are same-batch reruns under the same evaluation pipeline, and CaseRouter answers are assembled deterministically from the routed case in 3,637 of 3,646 items.

Method	Target hit	Slot F1	Char F1
Naive (dense chunks)	0.2660	0.2541	0.5168
LightRAG	0.3267	0.3041	0.5475
Hyper-RAG	0.4640	0.4315	0.6438
BM25 chunks (sparse)	0.5254	0.5000	0.6067
Cog-RAG	0.5448	0.5198	0.5973
CaseRouter	**0.8661**	**0.8215**	**0.9207**

6 EXPERIMENTS

6.1 SYSTEMS COMPARED AND THE REFERENCE SYSTEM

Table 3 compares a dense chunk retriever (BGE-M3 over the normalized chunks, top-k search, no case or graph structure), a sparse BM25-over-chunks retriever, two graph-augmented RAG systems (LightRAG (Guo et al., 2025) and Hyper-RAG (Feng et al., 2026), on the same case corpora under the same interface), Cog-RAG under its released `direct` pipeline with `cog-hybrid` retrieval (Hu et al., 2026), and the deterministic reference system. Every reported number is a macro average over the 17 books, and all systems use the same frozen files, instructions, inference stack, and parser (Section 4 gives the composition of both targets); closed-book and random-case controls appear in Table 6. The evaluation asks whether retrieval preserves the case template, whether the routed case can cross sources for herb grounding, and whether gains survive inspection of each answer with its evidence path.

The reference system, CASEROUTER, instantiates case-preserving routing so that the benchmark has a measured upper reference and a documented failure profile; we claim no novelty for it. It reads a typed case memory $\mathcal{M}_{\text{case}}$ and an indexed herb space linked by $\mathcal{B}_{\text{bridge}}$, scores candidate cases with a fixed linear function $\text{score}(c \mid q) = \sum_{f \in \mathcal{F}} w_f \, \phi_f(q, c)$ over match and similarity features, and reads the case memory by exact key lookup with a lexical BM25 fallback on misses. Every decision is a fixed, inspectable rule whose evidence route can be audited line by line; the feature set and the routing procedure are documented in the appendix, the exact constants are in the released repository, and all weights were frozen before the final comparison. The routed case then constrains the output slots, and answers are extracted rather than generated, so they vary only with the routed case and carry no decoding variance.

6.2 MAIN PATHOLOGY RESULTS

Table 3 shows CaseRouter recovering the disease-book case that underwrites a TCM answer: macro target hit 0.8661 versus 0.5448 for the strongest comparator Cog-RAG, improving target hit by $+0.3213$, slot F1 by $+0.3017$, and character F1 by $+0.3235$. Book-level bootstrap 95% confidence intervals for the target-hit and slot-F1 gains are 0.2508 to 0.3918 and 0.2587 to 0.3458, two-sided Wilcoxon signed-rank tests across the 17 books give $P = 1.53 \times 10^{-5}$, and all 17 books improve individually. The question-family stratification (per-family rates in the supplementary material) shows the improvement is not carried by one subtype: the largest gaps to the strongest comparator appear in modification-from-condition (0.0324 to 0.9823) and symptom-to-formula chain questions (0.1822 to 0.8988), the families most dependent on explicit case structure, and the smallest in formula selection (0.6652 to 0.8169). The anchor-versus-external split reduces single-book concerns: CaseRouter reaches target hit 0.9348 on the Spleen-Stomach anchor book and 0.8327 on the remaining 16 external books, against 0.3565 and 0.5565 for Cog-RAG (Table 4).

Table 5: Macro summary of the 17-book pathology ablation sweep. Challenge averages are computed only over books containing challenge questions.

Variant	Books	Target hit	Slot F1	Challenge hit	Challenge slot F1
CaseRouter	17	0.8383	0.7938	0.9879	0.7230
w/o typed case memory	17	0.8269	0.7873	0.8478	0.6255
w/o slot-aware realization	17	0.7787	0.7498	0.4571	0.4607
w/o modification routing	17	0.7634	0.7414	0.0000	0.0149

Table 6: Routing controls on the 17-book pathology benchmark (macro averages, deterministic slot realization; the reported configuration reaches 0.8661 in Table 3). Random replaces routing with a random same-book case, BM25 selects the case lexically, cross-case majority answers each slot by the majority across cases sharing the anchor, exact lookup is the router without the lexical fallback, and oracle supplies the gold case. The oracle's slot F1 (0.9462) reflects residual vocabulary-detection noise in the metric itself.

Configuration	Target hit	Slot F1	Char F1
Random case (same book)	0.0285	0.0272	0.2070
Closed-book (no retrieval)	0.1708	0.1609	0.4723
BM25 case matching	0.8149	0.7741	0.8751
Cross-case field majority	0.8144	0.7702	0.8613
Typed routing, exact lookup only	0.8383	0.7938	0.8928
CaseRouter (typed routing + lexical fallback)	**0.8656**	**0.8211**	**0.9203**
Oracle case (gold supplied)	0.9947	0.9462	0.9961

Table 4: Anchor-book and external-book pathology comparison. Anchor is Spleen-Stomach; external denotes the remaining 16 books.

Method	Anchor hit	Anchor slot F1	External hit	External slot F1
LightRAG	0.1435	0.1473	0.3381	0.3139
Hyper-RAG	0.3348	0.2997	0.4721	0.4397
Cog-RAG	0.3565	0.3897	0.5565	0.5280
CaseRouter	**0.9348**	**0.8601**	**0.8327**	**0.7901**

6.3 What the Benchmark Discriminates

The benchmark is built to isolate one capability, so we check that it discriminates between the mechanisms of a case-preserving system rather than rewarding one trick. Removing one mechanism at a time—typed case memory, slot-aware realization, and modification routing—changes the score in the expected order, with modification routing the strongest effect and typed case memory next, best supported on challenge-style questions. At the 17-book level (Table 5) challenge macro target hit falls from 0.9879 to 0.8478, 0.4571, and 0.0000; typed case memory has the smallest macro effect (-0.0114) because it concentrates in structure-dependent families, where removing it costs 0.1401. The three components jointly account for 0.1459, about 45% of the 0.3213 gap to Cog-RAG, with the rest arising from the full pipeline. Answers are copied from the routed case's field text in $3{,}637$ of $3{,}646$ items, so the metrics measure routing rather than generation.

Routing controls. Full-scale controls isolate routing while holding realization fixed (Table 6). Replacing routing with a random same-book case collapses macro target hit to 0.0285 and the closed-book model reaches 0.1708, while supplying the gold case raises it to 0.9947; the full system reaches 0.8656. Replacing typed routing with BM25 lexical matching (0.8149) or cross-case majority lookup (0.8144) costs 5.1 points, concentrated in books where the same anchors map to different slots. Structured case memory and slot-aware realization therefore carry the bulk of the performance, while typed routing adds the final margin where anchor-to-slot mappings are ambiguous.

Routing robustness. An exact-lookup router returns no candidate on a miss. Rewriting every benchmark question with a paraphrase model while preserving its anchors (3,146 of 3,646 questions) drops macro target hit from 0.8383 to 0.3838, with no candidate for 54.7% of paraphrased questions

versus 2.4% originally, while the oracle ($0.9947 \rightarrow 0.9935$) and random controls are unchanged; the loss therefore lies in routing, not in realization or evaluation. A lexical BM25 fallback on lookup misses recovers paraphrase stability ($0.3838 \rightarrow 0.8176$) and raises the original benchmark to 0.8656.

6.4 Error Taxonomy and Provenance

To locate difficulty rather than only score it, we classify every incorrect item of the reference system (490 of $3{,}646$, 13.4%) into four mechanical buckets. Routing selects the wrong case in 462 items (94.3% of errors); 26 items (0.7% of all items) have a gold alias that appears nowhere in that book's case text, which is a coverage limit rather than a system failure; realization accounts for 2; and no item fails because routing returns empty, since the lexical fallback absorbs that bucket. Difficulty here is therefore almost entirely case-routing difficulty, and annotation coverage is 99.3%. The stricter supported-answer rate, which requires target hit plus an evidence context hit, rises from 0.1617 (naive) and 0.5161 (Cog-RAG) to 0.7812 for CaseRouter (per-book values in the supplementary material).

6.5 Reverse Bridge: A Negative Result on Typed Field Scoring

Across 1,237 herb-name-masked questions, full-corpus clue matching reaches only 0.3290 weighted target hit, whereas CaseRouter reaches 0.9863 with a full-path answer rate of 0.9766; a raw-text control that scores the same routed candidate set against whole herb entries instead of typed herb fields reaches 0.9887. Once candidates are bounded by the routed case, typed field scoring adds no measurable benefit, so the gain comes entirely from candidate construction: formula-only candidates drop to 0.8278 (0.0000 on condition clues) while matched modification herbs recover 0.7647 and 0.9539 (paired exact McNemar $P = 2.27 \times 10^{-13}$; per-book tables in the supplementary material).

6.6 Discussion

Structure, not retrieval scale, sets the ceiling. Dense (0.2660), sparse (0.5254), generic graph retrieval (0.4640 for Hyper-RAG, 0.3267 for LightRAG) and the strongest comparator Cog-RAG (0.5448) all stay below 0.55 macro target hit, so generic graph structure over the same corpora adds little over sparse matching. What separates the reference system (0.8661) is that its answer is taken from a routed source field rather than assembled from retrieved spans, so the gap measures case preservation rather than ranking quality. Two further results sharpen the diagnosis: bounding candidates, not field typing, is what carries cross-source grounding, since typed-field scoring is indistinguishable from raw-text matching once candidates are fixed; and the controls turn a score into a diagnosis, since a paraphrase probe drops exact-key routing by 0.4545 while leaving the oracle intact, and the error taxonomy attributes 94.3% of failures to wrong-case routing rather than to realization or evaluation. The practical implication for system design is narrow: retrieving more relevant text does not close the gap unless a system can commit to one source case and return the field that case carries, and a system that cannot expose its route is penalized even when its final string is right.

7 Conclusion

We introduced CaseTCM, a benchmark and evaluation protocol for case-preserving cross-source retrieval in traditional Chinese medicine, in which gold answers are source fields and the evidence route is scored rather than assumed. A deterministic reference system reaches 0.8661 macro target hit against 0.5448 for the strongest baseline on all 17 books, and the benchmark's value lies in what it isolates: flat retrieval plateaus below 0.55, typed field scoring adds nothing once candidates are bounded, and exact-key routing is paraphrase-brittle until a lexical fallback is added. Extending the probe to a second corpus family, and to sources whose case templates are less structured than disease monographs, would test whether retrieval-time provenance fidelity can still be scored when the source scaffold must itself be inferred. The frozen benchmark and its scripts are released for reuse.

AI USE STATEMENT

We used LLM-based tools for drafting, editing, and formatting the manuscript, for automated consistency checks on experimental summaries, and for coding assistance during the implementation and evaluation pipeline. One diagnostic control in the protocol rewrites a subset of benchmark questions with a local open-weight LLM; these rewrites serve only as an evaluation perturbation, while all benchmark content, gold answers, and reported numbers come from the frozen source files and evaluation scripts rather than from any model. We manually reviewed all AI-assisted text, claims, citations, and code before submission, including cross-checking every reported number against the frozen benchmark, ablation, and provenance result files. We take responsibility for the final content of this work.

ETHICS STATEMENT

This study used OCR-processed literature materials only. No patients, identifiable personal data, biological specimens, intervention allocations, or animal experiments were involved.

REPRODUCIBILITY STATEMENT

The frozen benchmark, the evaluation protocol, the graph-construction code, and the routing weights are documented in the supplementary material and released as an anonymized repository at `https://anonymous.4open.science/r/Case-tcm-code`, which contains the frozen benchmark files used for every reported number. All compared systems use the same frozen benchmark files, short-answer instructions, shared local inference stack, and evaluation parser; the main rerun details and additional tables are provided in the supplementary material.

REFERENCES

Mohammad Alkhalaf, Ping Yu, Mengyang Yin, and Chao Deng. Applying generative ai with retrieval augmented generation to summarize and extract key clinical information from electronic health records. *Journal of biomedical informatics*, 156:104662, 2024.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the association for computational linguistics: ACL 2024*, pp. 2318–2335, 2024.

Vincent CH Chung, Robin ST Ho, Xinyin Wu, and Justin CY Wu. Incorporating traditional chinese medicine syndrome differentiation in randomized trials: methodological issues. *European Journal of Integrative Medicine*, 8(6):898–904, 2016.

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

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

Dennis Fast, Lisa C Adams, Felix Busch, Conor Fallon, Marc Huppertz, Robert Siepmann, Philipp Prucker, Nadine Bayerl, Daniel Truhn, Marcus Makowski, et al. Autonomous medical evaluation for guideline adherence of large language models. *NPJ Digital Medicine*, 7(1):358, 2024.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 3558–3565, 2019.

Yifan Feng, Hao Hu, Shihui Ying, Xingliang Hou, Shiquan Liu, Mingyuan Yang, Junchang Li, Shaoyi Du, Nanning Zheng, Han Hu, et al. Hyper-rag: Combating llm hallucinations using hypergraph-driven retrieval-augmented generation. *Nature Communications*, 2026.

Yue Gao, Zizhao Zhang, Haojie Lin, Xibin Zhao, Shaoyi Du, and Changqing Zou. Hypergraph learning: Methods and practices. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2548–2566, 2020.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tian Ao, and Chao Huang. Lightrag: Simple and fast retrieval-augmented generation. In *EMNLP (Findings)*, pp. 10746–10761, 2025.

Bernal J Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. *Advances in neural information processing systems*, 37:59532–59569, 2024.

Hao Hu, Yifan Feng, Ruoxue Li, Rundong Xue, Xingliang Hou, Zhiqiang Tian, Yue Gao, and Shaoyi Du. Cog-rag: cognitive-inspired dual-hypergraph with theme alignment retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31032–31040, 2026.

Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*, 40(Supplement_1):i119–i129, 2024.

Yu He Ke, Liyuan Jin, Kabilan Elangovan, Hairil Rizal Abdullah, Nan Liu, Alex Tiong Heng Sia, Chai Rick Soh, Joshua Yi Min Tung, Jasmine Chiat Ling Ong, Chang-Fu Kuo, et al. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. *npj Digital Medicine*, 8(1):187, 2025.

Jiyeong Kim, Michael L Chen, Shawheen J Rezaei, Tina Hernandez-Boussard, Jonathan H Chen, Fatima Rodriguez, Summer S Han, Rayhan A Lal, Sun H Kim, Chrysoula Dosiou, et al. Artificial intelligence tools in supporting healthcare professionals for tailored patient care. *NPJ Digital Medicine*, 8(1):210, 2025.

Simone Kresevic, Mauro Giuffrè, Milos Ajcevic, Agostino Accardo, Lory S Crocè, and Dennis L Shung. Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. *NPJ digital medicine*, 7(1):102, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, et al. Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset. *Advances in Neural Information Processing Systems*, 36:52430–52452, 2023.

Siru Liu, Allison B McCoy, and Adam Wright. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. *Journal of the American Medical Informatics Association*, 32(4):605–615, 2025a.

Yu Liu, Yishan Yuan, Keming Yan, Yuanyuan Li, Valeria Sacca, Sierra Hodges, Mattia Cannistra, Pauline Jeong, Jiani Wu, and Jian Kong. Evaluating the role of large language models in traditional chinese medicine diagnosis and treatment recommendations. *NPJ digital medicine*, 8(1): 466, 2025b.

Danping Pan, Yilei Guo, Yongfu Fan, and Haitong Wan. Development and application of traditional chinese medicine using ai machine learning and deep learning strategies. *The American journal of Chinese medicine*, 52(03):605–623, 2024.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. Kilt: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2523–2544, 2021.

Yaxuan Ren, Xufei Luo, Ye Wang, Haodong Li, Hairong Zhang, Zeming Li, Honghao Lai, Xuanlin Li, Long Ge, Janne Estill, et al. Large language models in traditional chinese medicine: a scoping review. *Journal of Evidence-Based Medicine*, 18(1):e12658, 2025.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, et al. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. *Advances in Neural Information Processing Systems*, 37:21999–22027, 2024.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature medicine*, 31(3):943–950, 2025.

Hirotaka Takita, Daijiro Kabata, Shannon L Walston, Hiroyuki Tatekawa, Kenichi Saito, Yasushi Tsujimoto, Yukio Miki, and Daiju Ueda. A systematic review and meta-analysis of diagnostic performance comparison between generative ai and physicians. *NPJ digital medicine*, 8(1):175, 2025.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*, 2021.

Chaoyu Wang, Guowei Dai, Yue Luo, Chuanbiao Wen, and Qingfeng Tang. Chinese medicine in the era of artificial intelligence: challenges and development prospects. *The American Journal of Chinese Medicine*, 53(02):353–384, 2025.

Dingqiao Wang, Jiangbo Liang, Jinguo Ye, Jingni Li, Jingpeng Li, Qikai Zhang, Qiuling Hu, Caineng Pan, Dongliang Wang, Zhong Liu, et al. Enhancement of the performance of large language models in diabetes education through retrieval-augmented generation: comparative study. *Journal of medical Internet research*, 26:e58041, 2024a.

Li Wang, Xi Chen, XiangWen Deng, Hao Wen, MingKe You, WeiZhi Liu, Qi Li, and Jian Li. Prompt engineering in consistency and reliability with the evidence-based guideline for llms. *NPJ digital medicine*, 7(1):41, 2024b.

Xidong Wang, Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, et al. Cmb: A comprehensive medical benchmark in chinese. In *Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers)*, pp. 6184–6205, 2024c.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6233–6251, 2024a.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. Improving retrieval-augmented generation in medicine with iterative follow-up questions. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pp. 199–214. World Scientific, 2024b.

Xinyi Xu, Huayu Xu, Yue Shang, Ran Zhu, Xiaoxu Hong, Zonghua Song, and Zhaopeng Yang. Development of the general chapters of the chinese pharmacopoeia 2020 edition: A review. *Journal of Pharmaceutical Analysis*, 11(4):398–404, 2021.

Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, et al. Almanac—retrieval-augmented language models for clinical medicine. *Nejm ai*, 1(2):AIoa2300068, 2024.

Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf. Learning with hypergraphs: Clustering, classification, and embedding. *Advances in neural information processing systems*, 19, 2006.

Supplementary Material for CaseTCM: A Benchmark for Case-Preserving Cross-Source Retrieval in Traditional Chinese Medicine

A Data Availability, Datasheet and Maintenance

A.1 Documentation and intended uses

CaseTCM is a retrieval-evaluation resource: it is intended for measuring whether a system recovers the source-native case template behind a TCM answer and whether that route can be inspected. It is not a clinical decision resource and should not be used to validate medical safety. Each book ships as line-delimited JSON files: `qa_benchmark.jsonl` (questions with gold answers, aliases, and gold slots), `cases.jsonl` (typed case records), `source_chunks.jsonl` or `case_source_chunks.jsonl` (retrieval corpus), `vocab.json` (slot vocabularies used by slot F1) and `stats.json`. All files are UTF-8 JSON Lines and are read with the loader in the released repository (Section A.2); no proprietary format is required.

A.2 Code and benchmark release

The benchmark files, the construction scripts, the shared evaluation parser, and the reference-system implementation are released together as an anonymized repository at `https://anonymous.4open.science/r/Case-tcm-code`. That repository contains the frozen benchmark files used for every reported number, so the results can be recomputed without any private resource beyond local model weights. A de-anonymized release with a persistent identifier will replace it after the review period.

A.3 Construction, hosting and maintenance

Construction is deterministic: OCR-normalized monographs are parsed into typed fields, questions are generated from those slots, and pharmacopoeia entries are linked through bridge rows; the scripts are in the released repository. The source books and the Chinese Pharmacopoeia are published literature; the derived questions and short evidence snippets are distributed under a CC BY license together with the code, and maintained by the authors, who will process corrections and rights requests.

A.4 Author responsibility

The authors confirm that they bear all responsibility in case of violation of rights and for the content of the released material, and that the derived artifacts (questions and short evidence snippets) are distributed under the license stated above.

B Benchmark Inventory and Additional Results

Benchmark construction and evaluation were frozen before the final comparison pass because monograph-derived TCM corpora contain OCR noise, formula aliases, and variable answer granularity. A single normalization and parsing pipeline therefore reduces the risk that apparent gains are driven by benchmark-specific answer handling rather than by retrieval structure.

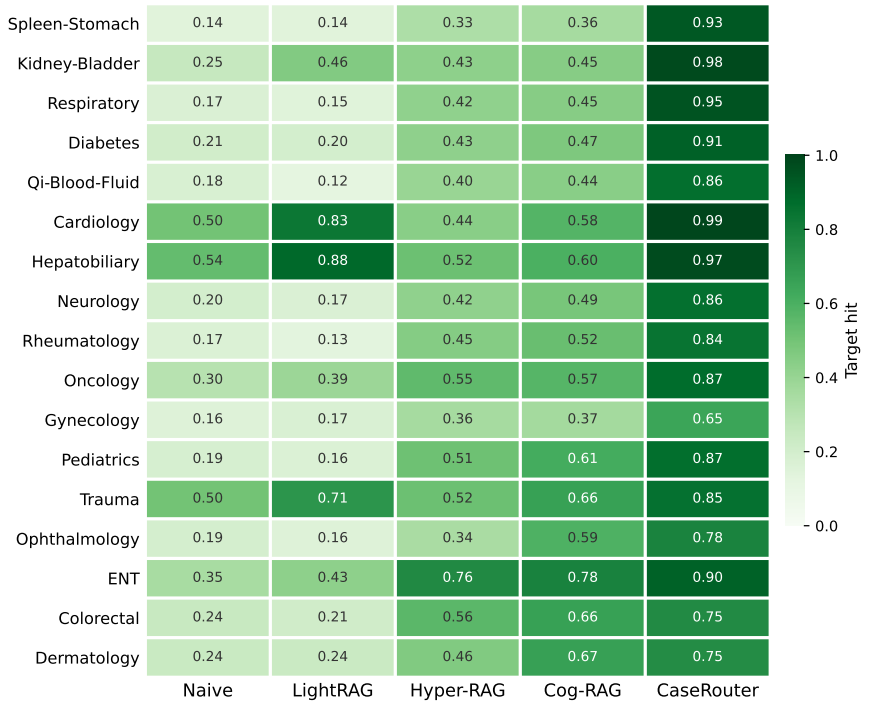

Figure 3: Per-book target hit across the five systems with per-book runs (naive, LightRAG, Hyper-RAG, Cog-RAG, CaseRouter); BM25-over-chunks is reported at the macro level only (Table 3 of the main paper). Disease books are ordered by the CaseRouter minus Cog-RAG gain.

Table 7: Inventory of the 17-book pathology benchmark.

Disease book	Source chunks	Case chunks	Cases	QA pairs
Ear-Nose-Throat	66	21	56	115
Pediatrics	254	46	124	267
Rheumatology	85	33	66	227
Gynecology	72	31	103	219
Hepatobiliary	161	11	19	112
Colorectal	170	35	62	136
Trauma	261	9	30	62
Respiratory	227	35	45	254
Neurology	191	60	99	435
Dermatology	192	33	71	153
Spleen-Stomach	39	39	39	230
Qi-Blood-Fluid	181	75	111	458
Kidney-Bladder	234	17	26	150
Diabetes	216	47	109	430
Cardiology	82	9	20	118
Ophthalmology	174	55	75	154
Oncology	94	17	59	126

The primary benchmark separates single-slot questions ($N = 3{,}060$), modification-condition questions ($N = 339$), and symptom-to-formula chain questions ($N = 247$). Individual books contribute 62 to 458 questions, spanning different syndrome vocabularies, formula naming conventions, modification density, and OCR quality. Question counts scale with recoverable structured slots rather than with corpus length: books with dense, well-structured case templates yield more questions per source chunk, while longer books with sparser structured content yield fewer. The five reverse bridge books contain 176 to 356 herb-name-masked questions, requiring traversal from disease-book pathology to pharmacopoeia evidence under a herb-name-masked formulation.

C Pathology Ablations and Provenance

Table 8: Focused pathology ablation on the Spleen-Stomach anchor benchmark. Overall $N = 230$; challenge questions are modification-from-condition prompts ($N = 35$).

Variant	Overall hit	Overall slot F1	Challenge hit	Challenge slot F1
CaseRouter	0.9348	0.8601	1.0000	0.6675
w/o typed case memory	0.8435	0.7861	0.9143	0.6251
w/o slot-aware realization	0.9043	0.8519	0.8000	0.6133
w/o modification routing	0.7826	0.7603	0.0000	0.0114

Table 9: Book-level pathology support audit. Supported-answer rate requires both target hit and gold-evidence context hit.

Disease book	Naive	Light	Hyper	Cog	Router
ENT	0.3043	0.3826	0.7565	0.7913	0.8870
Pediatrics	0.1386	0.1311	0.4794	0.5468	0.8427
Rheumatology	0.0969	0.1101	0.4097	0.5639	0.7753
Gynecology	0.1142	0.1507	0.3242	0.4384	0.6027
Hepatobiliary	0.3036	0.4643	0.3929	0.4196	0.7857
Colorectal	0.1838	0.1912	0.5368	0.6838	0.6912
Trauma	0.1452	0.4355	0.5161	0.5968	0.8387
Respiratory	0.0709	0.0866	0.2756	0.4370	0.7638
Neurology	0.1034	0.1218	0.3724	0.4920	0.7770
Dermatology	0.2288	0.2288	0.4641	0.6275	0.6993
Spleen-Stomach	0.0913	0.0957	0.2696	0.3000	0.7652
Qi-Blood-Fluid	0.0590	0.0852	0.3188	0.4017	0.7511
Kidney-Bladder	0.0600	0.1200	0.2667	0.3467	0.7933
Diabetes	0.1093	0.1209	0.3884	0.5047	0.8837
Cardiology	0.2627	0.4661	0.2881	0.5254	0.8220
Ophthalmology	0.1753	0.1558	0.3247	0.5584	0.7597
Oncology	0.3016	0.3810	0.5476	0.5397	0.8413
Macro average	0.1617	0.2193	0.4077	0.5161	0.7812

Each ablation removes one mechanism: removing typed case memory replaces the key-indexed retrieval with a generic similarity rerank over all cases; removing slot-aware realization freely generates the answer from the routed case context instead of copying its fields; removing modification routing disables the modification-condition filter. In the focused Spleen-Stomach ablation, removing typed case memory lowers overall target hit from 0.9348 to 0.8435 with a paired delta of 0.0913, bootstrap 95% CI 0.0478–0.1348, and exact McNemar $P = 1.04 \times 10^{-4}$. It lowers chain target hit from 1.0000 to 0.6667. Removing slot-aware realization lowers overall hit to 0.9043 and challenge hit from 1.0000 to 0.8000. Removing modification routing lowers overall hit to 0.7826 and challenge hit to 0.0000, with challenge slot F1 falling from 0.6675 to 0.0114. At the 17-book level, paired tests support the importance of slot-aware realization and modification routing, while typed case memory is best supported on challenge-style questions and in the focused probe.

The supported-answer rate is stricter than target hit because it requires both a correct final answer and retrieved evidence containing a gold supporting span. CaseRouter's routed-case audit further shows that macro gold-case hit is 0.8014 and gold-case answer rate is 0.7980; the close match with the supported-answer rate indicates that QA gains are accompanied by recovery of the inspectable case anchor itself.

C.1 Generic Graph Augmentation versus Case Preservation

The five-method comparison clarifies what CaseRouter adds beyond flat retrieval and generic graph augmentation. Under the matched pathology setting, Cog-RAG is the strongest structured comparator, Hyper-RAG is the second-strongest graph baseline, and Light-RAG is more variable across specialties. Generic graph structure therefore helps when reproduced under the same protocol, but it still leaves a substantial gap when the user-facing requirement is the preservation of a case-level route through syndrome, pathogenesis, therapy, and formula. The per-book view in Figure 3 shows that the ordering is not driven by one specialty: CaseRouter improves over Cog-RAG on all 17 books, while the smallest positive gains appear in books with more OCR and formula-aliasing difficulty.

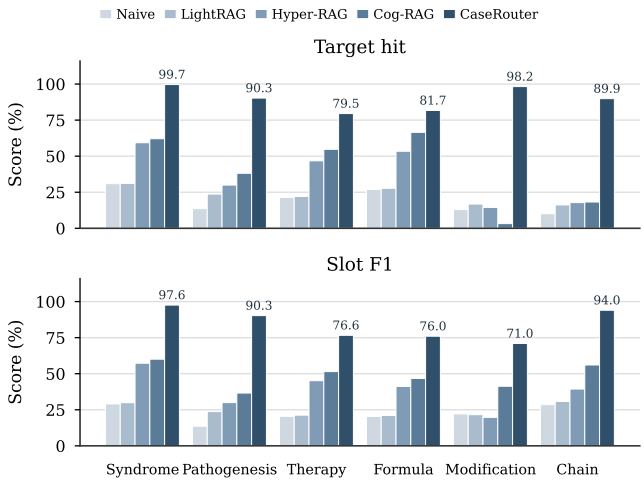

Figure 4: Pathology performance by question family (question-level rates). All five systems shown are same-batch reruns; BM25-over-chunks is reported at the macro level only. The ordering is stable across syndrome, pathogenesis, therapy, formula, modification, and chain questions, with the largest separations in modification and chain families.

The family-level decomposition is consistent with this reading: the separation is largest on the modification and chain families, which are the families that require a multi-field scaffold taken from one case, and smallest on families answerable from a single field. We do not attribute specific failure mechanisms to individual baselines; the comparison reports only what each system returns under the shared protocol.

D Reverse Bridge Controls and Extended Analysis

Table 10: Five-book herb-name-masked reverse benchmark. CaseRouter and the raw-text scoring control share the same routed pathology case and local candidate set; the control matches the clue against the full herb entry instead of typed herb fields. Clue-only matching scores the same clue over the full pharmacopoeia without candidate bounding.

Disease book	Questions	CaseRouter	Raw-text scoring	Clue-only
Spleen-Stomach	277	0.9747	0.9711	0.3285
Cardiology	176	0.9886	0.9943	0.3580
Respiratory	356	0.9944	0.9944	0.3455
Kidney-Bladder	249	0.9799	0.9920	0.2450
Hepatobiliary	179	0.9944	0.9944	0.3855
Weighted overall	1237	0.9863	0.9887	0.3290

Table 11: Reverse candidate-builder decomposition using the same raw-text clue matcher. All variants share the routed pathology case; they differ only in how the local herb candidate set is assembled. Overall $N = 1{,}237$; condition-clue $N = 204$.

Candidate builder	Overall hit	Condition-clue hit
Full routed candidate set (raw-text scoring)	0.9887	0.9755
Formula + matched modification herbs	0.9539	0.7647
Formula herbs only	0.8278	0.0000

The clue-only control scores the same pharmacopoeia clue over the full herb graph and remains at 0.3290 weighted target hit. The actual routed runtime has median local candidate size 8, range 1 to 37, and a weighted random-hit level of 0.1718. The candidate-builder decomposition shows that formula-only candidates remain adequate for formula-clue questions but fail on condition-clue

questions, whereas restoring modification-aware local candidates recovers most of the reverse performance. These controls indicate that the reverse task is not reducible to full-corpus clue matching; its difficulty is preserving the pathology-conditioned herb search space and resolving the local clue within it.

D.1 Automatic Metric Definitions

Let $\mathrm{norm}(\cdot)$ be the answer normalizer used by the shared evaluation parser: Unicode whitespace, punctuation, LaTeX markup, and non-text symbols are removed before matching. For an item i, let y_i denote the normalized model answer, C_i the normalized returned evidence context, and $G_i = \{a_{i,1}, \ldots, a_{i,K_i}\}$ the normalized set of gold aliases. Target hit is

$$H_i = \mathbb{1}\{\exists\, a \in G_i :\ a \text{ is a substring of } y_i\}\,, \tag{1}$$

and pooled target hit is $(1/N)\sum_{i=1}^{N} H_i$. Book-level macro target hit averages the per-book pooled rates first:

$$\mathrm{TH}_{\mathrm{macro}} = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{N_b}\sum_{i=1}^{N_b} H_i, \tag{2}$$

where B is the number of books and N_b the number of items in book b. The reverse-bridge tables use per-book weights $\mathrm{TH}_w = \sum_b N_b \overline{H}_b / \sum_b N_b$, which equals the question-level pooled rate when every question is assigned to exactly one book.

Character F1 is computed from character multiset overlap. Let $O(p, g)$ be the number of common normalized characters counted with multiplicity between predictions p and gold g. Then

$$F_{\mathrm{char},i} = \max_{a \in G_i} \frac{2\,O(y_i, a)}{|y_i| + |a|}, \tag{3}$$

and the reported character F1 is the item-level mean of $F_{\mathrm{char},i}$.

Slot F1 uses normalized vocabulary mentions detected in y_i. Let $G_{i,s}$ and $P_{i,s}$ be the gold and predicted normalized sets for slot s, and define $h_i = \sum_s |G_{i,s} \cap P_{i,s}|$, $p_i = \sum_s |P_{i,s}|$, and $g_i = \sum_s |G_{i,s}|$. The per-item slot F1 is

$$\mathrm{slotF1}_i = \frac{2h_i}{p_i + g_i}, \tag{4}$$

with $p_i = g_i = 0$ defined as 0.0; macro slot F1 is the item-level or book-level mean of $\mathrm{slotF1}_i$ following the same convention as Eq. (2).

Context hit is

$$E_i = \mathbb{1}\{\exists\, a \in G_i :\ a \text{ is a substring of } C_i\}\,, \tag{5}$$

and the supported-answer rate is $(1/N)\sum_i H_i E_i$. Gold-case hit is $\mathbb{1}\{\hat{c}_i \in G_i^{\mathrm{case}}\}$, where $\hat{c}_i$ is the selected case identifier and G_i^{case} is the gold case set; gold-case answer rate is the supported answer with $\hat{c}_i \in G_i^{\mathrm{case}}$. The reverse full-path answer rate is the stricter reverse counterpart that requires target hit plus evidence containing the routed case anchor, the matching bridge row, and the pharmacopoeia field used to construct the clue. All parsers and aggregate scripts are included in the released repository (Section A.2).

E Detailed Method and Evaluation Protocol

E.1 Corpus and Schema

The pathology schema includes disease, syndrome, pathogenesis, symptom, sign, tongue quality, tongue coating, pulse, therapy, treatment principle, formula, herb, dosage, contraindication, caution, and modification-condition nodes, with lighter use of disease-location and constitution cues where recoverable. Answer normalization was treated as part of benchmark definition: formula aliases, herb naming variants, punctuation differences, and short OCR artifacts were mapped to canonical answer forms where equivalence was explicit, while original short-answer surface forms were retained for faithful book-style output evaluation.

Algorithm 1 Pathology QA with case-grounded slot realization

Require: query q, $\mathcal{G}_{\text{path}}$, $\mathcal{M}_{\text{case}}$

1: detect requested slot type or chain format from q
2: retrieve candidate cases from $\mathcal{M}_{\text{case}}$ by exact key lookup on symptom, syndrome, therapy, formula, and condition keys
3: if no candidate remains, rank all cases by lexical BM25 over case text and take the best
4: score each candidate by $\text{score}(c \mid q)$ and select the top-ranked case $c^\star$
5: retrieve supporting pathology evidence from $c^\star$
6: copy the requested slot or ordered chain from $c^\star$ (generation only if the copy is empty)
7: return final answer together with supporting pathology evidence

Algorithm 2 Bridge QA with pathology-to-pharmacopoeia traversal

Require: query q, routed case $c^\star$, $\mathcal{B}_{\text{bridge}}$, $\mathcal{G}_{\text{pharm}}$

1: identify formula, syndrome, and modification-condition anchors from $c^\star$
2: collect bridge candidates consistent with these anchors
3: score each bridge row by $\text{score}(b \mid q, c^\star)$
4: select the top-ranked herb within the routed local candidate set
5: retrieve efficacy, indication, dosage, and caution evidence from $\mathcal{G}_{\text{pharm}}$
6: aggregate bridge-consistent pharmacopoeia evidence
7: return the selected herb together with its aggregated evidence

E.2 Pathology Case Routing and Realization

Operational routing uses a staged deterministic filter before context reranking. Symptom and chain questions first use exact symptom keys; other pathology questions use normalized syndrome keys with formula-name fallback when needed. Formula mentions restrict the candidate pool to cases whose normalized formula aliases match the query. Herb-efficacy questions retain only cases containing the queried herb, while modification and herb-indication questions additionally require an exact or high-scoring condition match before reranking. When more than one candidate remains, a fixed context-reranking score combines exact key matches, containment bonuses on the case fields, and lexical similarity between the query and the case excerpt, with additional weight on the fields the question asks about; the feature set, the score form, and the constants are in the released repository. These constants were fixed before the final benchmark comparison and were not tuned on the benchmark.

E.3 Bridge Scoring and Rerun Details

On the bridge side, candidate herbs are first restricted to the herbs recoverable from the routed formula text or from the matched modification clause of the selected pathology case. Within this local pool, the final herb is selected by an exact clue-match bonus and a character-overlap similarity term on the requested pharmacopoeia field, with the clue-match bonus bounded so that a long verbatim clue cannot dominate the comparison; the exact form is in the released repository. In the reported reverse reruns, model fallback was disabled and the herb-name-masked reruns were regenerated with benchmark-only herb labels excluded from case routing, so that no label occurring only in the benchmark can enter routing. The five bridge books share the same disease-book case files, pharmacopoeia bridge corpora, and clue-construction protocol.

E.4 Comparator and Implementation

Cog-RAG baselines were reproduced with the released direct pipeline using `cog-hybrid` retrieval and the local Qwen3.5-4B instruction model with BGE-M3 embeddings (Chen et al., 2024). The main pathology comparison uses a local Qwen3.5-4B generator across naive retrieval, Cog-RAG, Hyper-RAG, LightRAG, and CaseRouter, with the same short-answer instruction. No parameter fine-tuning was used. Hyper-RAG and LightRAG are run on the same disease-book case corpora with the shared local inference stack and embedding backend. All systems share the frozen benchmark files, unified answer templates, and evaluation parser. Baselines were reproduced from their official released code under this matched local setting; their numbers therefore reflect the shared 4B generator and this corpus rather than the original papers' API-scale configurations, and should not be

read as the originals' best attainable performance. All local runs use greedy decoding (sampling disabled) with a 256-token output budget and float16 weights; decoding parameters were fixed before the benchmark comparison. Within CaseRouter, answers are realized deterministically by copying the routed case's fields, and the instruction model is invoked only when slot filling returns no answer (9 of 3,646 items in the reported runs). Replacing the 4B model with Qwen3.5-2B reproduces target hit 0.8661 and slot F1 0.8215 and moves character F1 by 0.0002 (0.9207 to 0.9209).

E.5 Data Availability and Licensing

The 17 disease books and the Chinese Pharmacopoeia are publicly published literature. The benchmark releases derived questions and short evidence snippets for research evaluation; users are responsible for checking redistribution terms in their jurisdiction before reuse.

E.6 Statistical Analyses

The main 17-book pathology comparison treats each disease book as one paired unit and uses two-sided Wilcoxon signed-rank tests across books. The 17-book ablation sweep uses one-sided Wilcoxon signed-rank tests with the alternative that the full model outperforms the ablated variant. Focused anchor-book ablations and the five-book reverse bridge benchmark report paired bootstrap 95% confidence intervals over question-level deltas together with exact McNemar tests on paired target-hit outcomes. Reverse-bridge bootstrap resamples questions; questions sharing a routed case are correlated, so these intervals should be read as question-level rather than case-level uncertainty.