Toward Trustworthy Multi-Agent Clinical Decision Support: Dual-RAG with Conditional Reviewer Integration

Wangshu Zhu [1]

Abstract

We present Dual-RAG with Conditional Reviewer Integration, a multi-agent framework for clinical decision support that bridges structured biomedical knowledge and experiential case-based reasoning. Our approach introduces three innovations: (i) a dual-path retrieval strategy that mirrors real physician diagnostic workflows, (ii) a *Reception Doctor* mechanism for symptom extraction and interpretable reasoning decomposition, and (iii) a conditionally triggered Reviewer agent that calibrates final decisions without the performance degradation of naive reviewer setups. Furthermore, the system incorporates a self-reinforcing case memory, automatically assimilating new encounters into its retrieval base, thereby achieving continuous performance improvement over time. Evaluated on two diagnostic benchmarks, our framework consistently outperforms single-agent and baseline RAG approaches, achieving state-of-the-art accuracy while maintaining interpretability and auditability. These results demonstrate the promise of combining structured knowledge, experiential learning, and conditional oversight to advance trustworthy clinical LLM deployment in real-world practice.

1. Introduction

Large language models (LLMs) are increasingly being considered for clinical decision support, where physicians must synthesize complex patient data, guidelines, and experiential knowledge to reach trustworthy conclusions. Despite impressive advances in reasoning capability, current clinical LLM systems face two critical challenges. First, their decision-making process is often opaque, limiting physician trust and hindering adoption in high-stakes environments. Second, attempts to enhance transparency or interpretability often come at the cost of degraded predictive accuracy, creating a fundamental tradeoff between trustworthiness and performance.

Existing approaches to trustworthy AI in healthcare tend to fall into two categories. On one hand, retrieval-augmented generation (RAG) pipelines improve grounding by linking LLM outputs to external knowledge, yet they typically rely on a single knowledge source and struggle to represent the nuanced way clinicians integrate both structured guidelines and case-based experiential reasoning. On the other hand, multi-agent debate frameworks provide a mechanism for consensus-building and error correction, but they frequently lack adaptive mechanisms to avoid unnecessary interventions that reduce overall accuracy. Thus, the field lacks a unified framework that reconciles accuracy with calibration, transparency, and clinical realism.

In this work, we propose a novel framework for multi-agent clinical decision support, termed *Dual-RAG with Conditional Reviewer Integration*. Our contributions are threefold. First, we design a dual-path retrieval strategy that mirrors real physician reasoning: one path retrieves structured biomedical knowledge (e.g., UMLS, clinical guidelines), while the other retrieves experience-based case databases. Second, we introduce a "TeachCoT" mechanism for symptom extraction and reasoning decomposition, enabling the system to generate explanations in a style similar to medical trainees reporting a case. Third, we augment a multi-agent debate with a Reviewer agent that is conditionally activated only under uncertainty or disagreement. This *conditional reviewer* avoids the performance degradation commonly observed in naive reviewer setups, while maintaining calibration and auditability of the final decision.

By bridging structured and experiential knowledge with a calibrated multi-agent workflow, our approach aims to deliver both high predictive accuracy and enhanced trustworthiness. We demonstrate that this framework preserves state-of-the-art accuracy while providing a transparent and auditable reasoning process, offering a practical path toward trustworthy clinical LLM deployment in real-world settings.

2. System Overview

Figure 1 illustrates our proposed workflow, **Dual-RAG with Conditional Reviewer Integration**, a multi-agent pipeline for trustworthy clinical decision support.

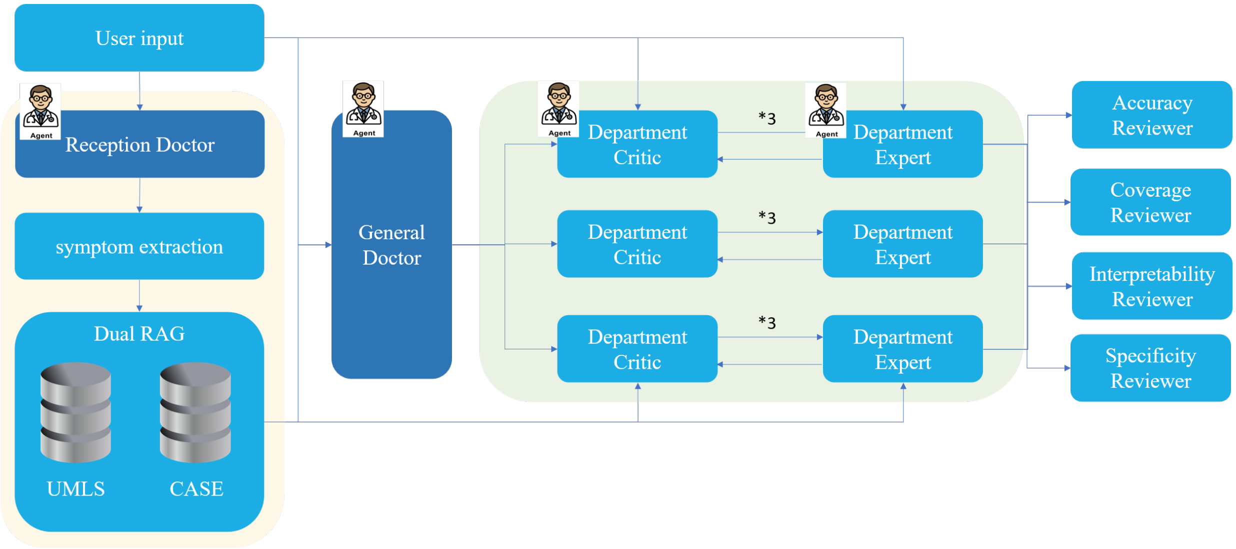

Figure 1. **Overview of our system.**

The pipeline begins with a *Reception Doctor* agent that collects user input (patient descriptions or EMR snippets) and extracts relevant symptoms. These features are then processed by the **Dual-RAG** module, which retrieves evidence from both structured biomedical terminologies (e.g., UMLS) and case-based repositories (e.g., historical patient records), ensuring guideline-grounded and experience-driven coverage. A *General Doctor* agent integrates the retrieved knowledge into a preliminary diagnostic reasoning chain.

To strengthen robustness, the system launches multiple *Department Critic–Expert* pairs, where critics challenge and experts defend reasoning across clinical specialties. Finally, when divergence or uncertainty remains, **Conditional Reviewers** evaluate the decision support output along four dimensions: accuracy, coverage, interpretability, and specificity. Only when these reviewers are satisfied is the final decision released.

2.1. Dual-RAG Knowledge Retrieval

A central component of our system is **Dual-RAG**, designed to emulate how real clinicians combine authoritative guidelines with experiential knowledge when reasoning about uncertain cases. Conventional retrieval-augmented generation (RAG) pipelines rely on a single evidence source, which often limits coverage or biases reasoning. In contrast, Dual-RAG establishes two complementary retrieval pathways:

Structured Knowledge Pathway. Leveraging curated biomedical ontologies such as UMLS and clinical guidelines, this pathway provides standardized, high-precision mappings between symptoms, diseases, and treatment options. It ensures that the reasoning process remains aligned with medical consensus.

Case-Based Pathway. In parallel, the system retrieves from a repository of historical patient records and annotated cases. This pathway contributes contextual richness and real-world variability, offering analogical reasoning beyond the rigidity of guidelines.

By fusing both pathways, Dual-RAG reduces the risk of under-coverage (guideline-only) or overfitting to spurious correlations (case-only), while more faithfully simulating how clinicians simultaneously "*check the guideline*" and "*recall similar patients*."

2.2. Reception Doctor Symptom Extraction

Accurate symptom extraction is essential for downstream reasoning, yet conventional rule-based or entity-tagging approaches often fail to capture temporal progression and patient-specific risk factors. In clinical practice, however, physicians rely heavily on structured narratives that emphasize *how* symptoms evolved and which contextual factors modulate risk.

To bridge this gap, we propose **Reception Doctor**, an instruction fine-tuned prompting strategy that trains the model to emulate the reporting style of medical trainees. Rather than returning isolated keywords, Reception Doctor produces hierarchically organized outputs that resemble a clinical history presentation: (i)Symptom onset and progression (ii)Aggravating and alleviating factors (iii)Relevant comorbidities and risk background

Reception Doctor Extraction Showcase

Input:
”user text”:”I have a rash on my face, especially around my nose. It’s red and painful, and there are some yellow sores that are oozing.”
”disease”:”impetigo”
Output:
”symptoms”:[”red painful facial rash”,”rash around nose”,”yellow oozing sores”],
”reasoning”:[”Why this fits impetigo: the presence of painful, red rash with yellow oozing sores around the nose is characteristic of impetigo, a superficial bacterial skin infection as described by the patient.”,
”differentials: herpes simplex (sores but usually clear fluid and less pain), contact dermatitis (red rash but less oozing and pain).”],

This emulation of clinical case reporting improves interpretability for both human reviewers and downstream multi-agent reasoning, while ensuring that extracted features are clinically meaningful.

2.3. Multi-Agent Debate

After the General Doctor produces up to three candidates via Dual-RAG, each candidate enters a focused, single-diagnosis debate between a Department *Critic* and a Department *Expert*. The goal is to *tighten* or *disconfirm* that specific candidate by interrogating *lever symptoms* and enforcing evidence-grounded answers, without introducing new diagnoses.

- *General Doctor (once):* Emits $\leq$3 candidates with initial confidence, reasoning, and labeled citations using CaseRAG (primary) and UMLS (auxiliary).
- *Department Critic (per candidate):* Identifies uncovered *lever symptoms* and asks questions (Q#) each round; labels prior answers as *fixed / not_fixed / cannot_fix*. Must not propose new diagnoses.
- *Department Expert (per candidate):* Answers each Q# with explicit evidence tags `[History RAG]` / `[Knowledge RAG]`; may adjust the *candidate’s own* confidence when warranted by evidence.

2.4. Conditional Reviewer Integration

Prior work often appends a “Reviewer” agent to re-evaluate all candidate outputs. While this may improve caution, it typically drives the system toward *over-conservatism*: the Reviewer down-scores otherwise correct answers, reducing overall accuracy. We propose a *gated Reviewer* that activates only under specific conditions:

$$\text{ReviewerActive} = \mathbb{1}\Big(|p_1 - p_2| < \delta \;\;\vee\;\; p_{\max} < \tau\Big),$$

Multi-Agent Debate Showcase

Patient input: *”I have sores on my face, mostly near my nose and lips. They are painful and have a yellow or rust-colored discharge.”*
GeneralDoctor:
Department: Dermatology
Diagnosis: Impetigo
Confidence: 0.6
Reasoning:
1) Patient has painful, red crusty facial sores with yellow or rust-colored discharge around nose and lips, typical for impetigo.
2) History RAG score 0.778 supports impetigo diagnosis based on described symptoms.
3) Impetigo is a superficial bacterial infection with characteristic facial lesions.
Reference:
History RAG: sores on my face, red and crusty, mostly around my nose and lips, painful, yellow or rust-colored fluid coming out.

Round 1 Critic:
Critics:
Q1: Does the patient have delayed healing of the sores?
Q2: Is there dryness or flakiness of the skin surrounding the affected areas?
Reviews:
A1: not fixed
A2: not fixed
Round 1 Expert:
Responds:
A1: Yes, the patient reports that the sores are “taking a long time to heal,” indicating delayed healing of the sores.
A2: I do not have evidence regarding dryness or flakiness of the skin surrounding the affected areas.
References:
History RAG: disease=impetigo — Key symptoms include ‘taking a long time to heal’ for sores around the nose, indicative of delayed healing.

Round 2 Critic:
Critics:
Q3: Does the patient have fever and systemic symptoms such as headache or malaise?
Reviews:
A1: fixed
A2: cannot fixed
A3: not fixed
Round 2 Expert:
Responds:
A3: Yes, the patient reports having fever, headache, malaise, and fatigue alongside the skin symptoms, indicating systemic involvement consistent with impetigo.
References:
History RAG: disease=impetigo — Evidence includes ‘fever, headache, and a rash on my arms and face’, ‘feeling really tired and I have a fever’, ‘fever and a headache’, ‘really struggling’, and ‘malaise’ reported by patient along with characteristic sores and rash.

Round 3 Critic:
Critics:
none
Reviews:
A1: fixed
A2: cannot fixed
A3: fixed

where p_1, p_2 are the top-2 candidate confidences, δ controls disagreement tolerance, and τ is a minimum confidence threshold.

When triggered, the Reviewer performs a higher-order consistency check across (i) Critic–Expert debate transcripts, (ii) RAG evidence tags, and (iii) known risk-symptom alignments from the knowledge base. Otherwise, the pipeline bypasses the Reviewer to preserve performance.

In the impetigo case (Sec. 2.3), the debate converged to $p = 0.90$, comfortably above τ, so the Reviewer remained inactive. By contrast, in cases where General Doctor candidates cluster around $p = 0.40$–0.55 with conflicting symptom matches (e.g., differentiating *eczema* vs. *contact dermatitis*), the Reviewer is triggered. It explicitly contrasts lever symptoms (e.g., exudative discharge vs. dry plaques) and applies meta-reasoning: if neither candidate substantiates required features, it withholds final endorsement.

This conditional integration maintains the pipeline's diagnostic sharpness while providing a transparent, second-layer safeguard when uncertainty or disagreement truly exists. The result is a system that is not only high-performing but also trustworthy in its restraint.

3. Results and Discussion

3.1. Experimental Setup

We evaluated our framework on two test sets: **symptom2disease** (single symptom input) and **conversation2disease** (multi-turn dialogue input). For both datasets, we split diseases by class into 80% for historical cases and 20% for held-out testing. The historical cases were processed by our agent pipeline and stored as the retrieval base. Baselines do not use retrieval; they simply output three ranked disease candidates from the SingleAgent. We report Top-1/2/3 accuracy and mean reciprocal rank (MRR).

3.2. Baseline Comparisons

Symptom2disease. As shown in Table 1, the SingleAgent baseline yields moderate performance (Top-1 ∼45%), which is slightly improved by CoT reasoning. The original RAG variant (direct case retrieval without agent-based processing) significantly improves performance when combined with CoT, reaching ∼75% Top-1 accuracy. This highlights the importance of incorporating prior cases when only symptoms are provided.

Conversation2disease. On the dialogue test set (Table 2), the baseline already achieves strong Top-1 performance (∼74%). CoT further boosts accuracy, and CoT+RAG reaches nearly 87% Top-1. This suggests that the richer context in conversations provides a stronger signal, with retrieval offering additional but smaller gains. We omit older multi-agent variants as they were deprecated during system iteration.

Table 1. Results on **symptom2disease** test set (20% holdout).

Method	Top-1	Top-2	Top-3	MRR
SingleAgent	45.28%	14.15%	6.13%	54.4%
SingleAgent+CoT	46.23%	9.91%	0.66%	53.38%
SingleAgent+CoT+RAG	74.53%	6.60%	0.47%	77.99%

Table 2. Results on **conversation2disease** test set (20% holdout).

Method	Top-1	Top-2	Top-3	MRR
SingleAgent	73.96%	11.98%	0.26%	80.82%
SingleAgent+CoT	77.08%	9.38%	2.08%	82.47%
SingleAgent+CoT+RAG	86.98%	5.73%	0.52%	90.02%

3.3. Results on Our Final System

Due to resource constraints, our final system was fully evaluated on 113 out of 212 cases from the **symptom2disease** set. We report both *doctor-first rerank* and *final reviewer decision*. Doctor-first reranking achieves near-perfect accuracy (Top-1 ∼96%, MRR ∼0.98). Naively applying the reviewer on all cases lowers Top-1 to ∼89%, indicating that unconditional reviewing can be overly conservative. Therefore, we adopt a **conditional reviewer trigger**, which maintains high performance while ensuring safety in uncertain cases.

3.4. Discussion

These results lead to several key observations:

- **Retrieval is essential for symptom-only inputs.** Without retrieval, accuracy stagnates at ∼45%. Case-based memory raises Top-1 by nearly 30 points.
- **Dialogue input reduces reliance on retrieval.** Conversations provide richer context, explaining why baselines are already strong.
- **Reviewer must be conditional.** Unconditional reviewing hurts performance. Our gating mechanism ensures high accuracy while preventing unsafe outputs in ambiguous cases.
- **Clinical implication.** Our framework achieves >95% Top-1 accuracy in single-symptom settings, marking a significant step toward practical, case-based decision support.

4. Conclusion and Future Work

In this work, we presented **Dual-RAG with Conditional Reviewer Integration**, a multi-agent clinical decision support framework that couples structured biomedical knowledge with experiential case retrieval. Our contributions are threefold: (i) a dual-path retrieval strategy mirroring physician reasoning, (ii) a TeachCoT mechanism that decomposes reasoning into interpretable trainee-style reports, and (iii) a conditional Reviewer agent that activates only under uncertainty or disagreement, avoiding the over-conservatism of naive reviewer pipelines.

A distinctive feature of our system is its ability to perform self-reinforcing case accumulation: each newly encountered case is automatically processed by agents and incorporated into the case-based RAG memory. This creates a continuously expanding, self-updating repository that improves diagnostic performance over time without requiring manual curation.

Looking ahead, we envision several directions for extending this work. First, broadening retrieval to include multimodal signals (e.g., lab tests, imaging) may further align system reasoning with real-world diagnostic practice. Second, refining agent specialization—for example, differentiating sub-agents for dermatology versus infectious disease—may enhance both accuracy and interpretability. Finally, prospective evaluation with clinicians will be critical to validate safety and utility in live clinical workflows.

By integrating structured knowledge, experiential memory, and conditional oversight, our framework demonstrates a practical path toward accurate, transparent, and continuously improving clinical LLM systems.

Acknowledgment

This work was conducted during an internship at **Graphen**, and I sincerely thank my mentors and colleagues for their valuable guidance and constructive feedback throughout the project. We also acknowledge the support and resources provided by the company, which made this research possible.