DIAGMEM: CLOSING THE DIAGNOSTIC MEMORY GAP FOR EVIDENCE-GROUNDED MULTI-AGENT RARE DISEASE REASONING

Anonymous Submission

ABSTRACT

Rare disease diagnosis often depends on evidence scattered across symptoms, laboratory results, family history, phenotypes and external disease knowledge. Current LLM-based diagnostic systems usually treat this process as prompt-level prediction. They may produce plausible disease rankings, but they do not maintain what has been established, which hypotheses have been considered, what remains uncertain or how new evidence changes the case. We introduce Patient Diagnostic Memory, a shared-memory framework for multi-agent rare disease diagnosis. For each patient, the framework maintains a diagnostic state, event records and semantic memory entries that store reusable clinical evidence. Specialized agents extract clinical facts, normalize phenotypes with Human Phenotype Ontology terms, interpret laboratory findings, analyze family history and retrieve external disease knowledge. A diagnostic controller reads the evolving memory to decide which evidence source to use next and when to produce a ranked diagnosis. We evaluate the framework on MIMIC-IV-Note and RareArena with four backbone LLMs. Memory-augmented diagnosis improves top-ranked accuracy across all models, with larger gains on rare disease cases. Average top-1 accuracy increases by 14.5 percentage points on RareArena and 4.3 points on MIMIC-IV-Note. These results suggest that patient-level diagnostic memory can make multi-agent clinical reasoning more traceable, modular and evidence-grounded.

1 INTRODUCTION

Rare disease diagnosis is challenging because evidence is sparse, fragmented, and distributed across symptoms, laboratory findings, family history, phenotypes, and external medical knowledge Faye et al. (2024); Benito-Lozano et al. (2022); Wu et al. (2025). Unlike ordinary disease classification Chen et al. (2024), it requires maintaining differential diagnoses, integrating weak evidence, and continuously revising hypotheses as new information emerges.

Current LLM-based diagnostic systems mainly perform prompt-level prediction by mapping patient records to ranked disease lists Singhal et al. (2025); McDuff et al. (2025). However, they lack explicit mechanisms for maintaining diagnostic state, evidence provenance, uncertainty, and hypothesis evolution. This limitation is critical for rare diseases, where reasoning requires combining standardized phenotypes, external knowledge, and comparisons among plausible alternatives rather than relying only on learned patterns Zhu et al. (2025); Zheng et al. (2025).

Multi-agent systems improve diagnostic reasoning by decomposing tasks into specialized agents Zhao et al. (2026); Köhler et al. (2009). Nevertheless, without shared patient-level memory, intermediate evidence remains transient and difficult to trace, reuse, and integrate across agents. We refer to this limitation as the *diagnostic memory gap*.

To address this gap, we introduce *Patient Diagnostic Memory*, a shared-memory framework for stateful multi-agent rare disease diagnosis. The framework maintains patient-specific diagnostic states, event records, and semantic memories, enabling persistent evidence accumulation and coordinated reasoning among agents.

We evaluate Patient Diagnostic Memory on MIMIC-IV-Note Johnson et al. (2023) and RareArena THUMedInfo (2025) using four backbone LLMs. The framework improves ranked di-

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.

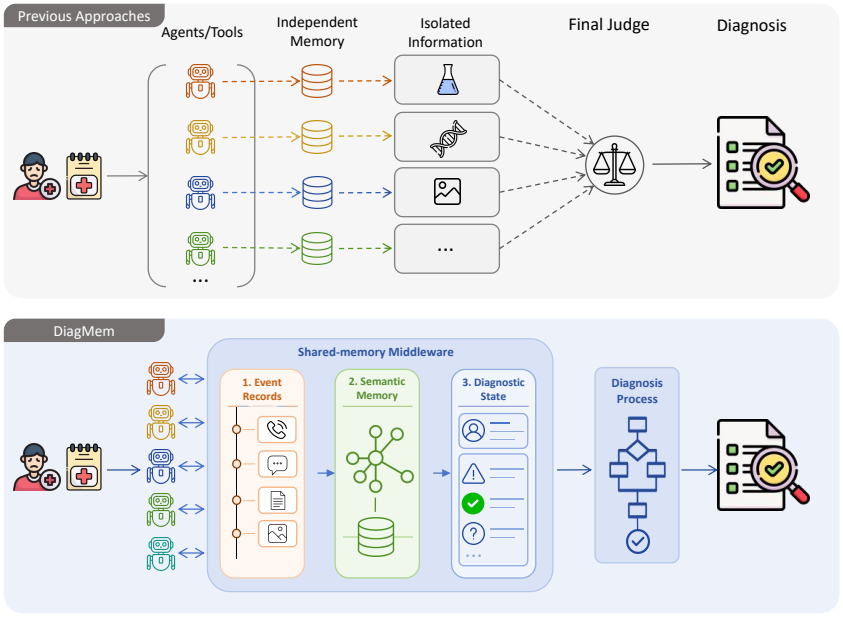

Figure 1: Comparison of previous multi-agent diagnosis approaches with our framework. The shared-memory middleware centralizes event records, semantic memories, and diagnostic state, enabling agents to collaboratively integrate evidence and produce a final ranked diagnosis.

agnosis accuracy across both datasets, with average R@1 gains of 4.3 percentage points on MIMIC-IV-Note and 14.5 points on RareArena.

This paper makes three contributions:

- We identify the diagnostic memory gap in multi-agent rare disease diagnosis and formulate diagnosis as maintaining patient-level state over evidence, hypotheses, and uncertainty.
- We introduce Patient Diagnostic Memory, a memory framework supporting persistent diagnostic state, event tracking, semantic clinical memory, and traceable reasoning.
- We demonstrate across two datasets and four LLMs that memory-centered diagnosis improves ranked diagnostic accuracy, particularly for rare disease cases.

2 Related Work

LLMs for Clinical and Rare Disease Reasoning. Recent LLMs have shown strong potential in diagnosis generation, medical question answering, and clinical note understanding Lan et al. (2025); McDuff et al. (2025); Singhal et al. (2025). Rare disease diagnosis is particularly challenging due to data sparsity, heterogeneous evidence, and the need to integrate symptoms, phenotypes, laboratory findings, family history, and external medical knowledge Chen et al. (2024); Köhler et al. (2009). However, most existing LLM-based diagnostic systems formulate diagnosis as single-pass or prompt-level prediction, making it difficult to explicitly track intermediate evidence, uncertainty, and evolving hypotheses across multiple reasoning steps.

Tool-Augmented, Retrieval-Augmented, and Multi-Agent LLMs. Tool-augmented and retrieval-augmented generation methods allow LLMs to query external knowledge bases, guidelines, or structured medical resources Lewis et al. (2020); Karpukhin et al. (2020); Lu et al. (2024). Meanwhile, multi-agent systems decompose complex tasks into specialized roles such as information extraction, knowledge retrieval, and decision making Zhao et al. (2026); Li et al. (2023); Tang et al. (2024), including recent medical diagnosis workflows Li et al. (2025). Although these methods improve knowledge access and modularity, tool calls and agent outputs are often treated as isolated prompt-level interactions Yao et al. (2023), rather than being integrated into a persistent patient-level diagnostic state.

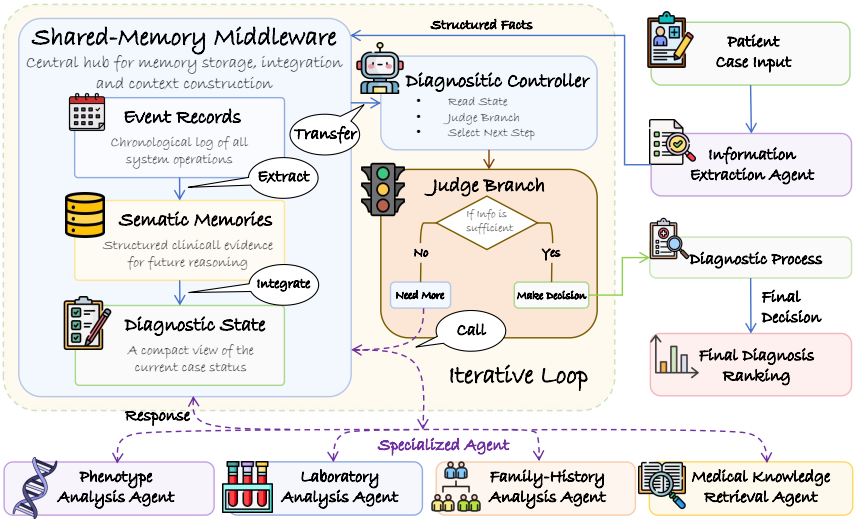

Figure 2: Multi-agent diagnostic framework. Patient data is extracted and stored in the Shared-Memory Middleware, while specialized agents provide evidence and the Diagnostic Controller iteratively produces a ranked list of candidate diagnoses.

Memory-Augmented LLMs and Our Contribution. Memory-augmented LLMs extend models with long-term memory, external stores, or persistent interaction histories Packer et al. (2023); Zhong et al. (2023); Park et al. (2023). These mechanisms support task continuity, but they are usually general-purpose and do not explicitly model clinical reasoning structures such as diagnostic states, evidence types, conflicts, provenance, and hypothesis evolution. In contrast, we introduce *Patient Diagnostic Memory*, a shared-memory framework that combines multi-agent specialization with patient-level diagnostic state, event records, and semantic clinical memories, enabling iterative, traceable, and evidence-grounded rare disease diagnosis beyond prompt-level prediction.

3 METHOD

We propose a middleware-centered multi-agent framework for rare disease diagnosis. The key idea is a shared-memory middleware that coordinates specialized agents and maintains patient-specific diagnostic memory.

For each patient case, the system maintains a diagnostic state:

$$S_t = \{F_t, H_t, U_t, D_t, K_t, E_t\} \tag{1}$$

where F_t denotes clinical facts, H_t candidate diagnoses, U_t unresolved questions, D_t controller decisions, K_t evidence conflicts, and E_t accumulated evidence.

Instead of directly predicting diagnosis from raw records, the framework decomposes reasoning into evidence-oriented agents coordinated by a diagnostic controller:

$$a_t = \pi(S_t) \tag{2}$$

where $\pi(\cdot)$ determines whether to invoke an agent or produce a diagnosis.

3.1 SHARED-MEMORY MIDDLEWARE

The shared-memory middleware provides a unified mechanism for storing, updating, and retrieving diagnostic information. Each patient maintains an independent memory instance.

Memory Structure. The patient-specific memory is defined as:

$$\mathcal{M}_t = (S_t, \mathcal{E}_t, \mathcal{R}_t) \tag{3}$$

where S_t, $\mathcal{E}_t$, and $\mathcal{R}_t$ denote the diagnostic state, event records, and semantic memories, respectively.

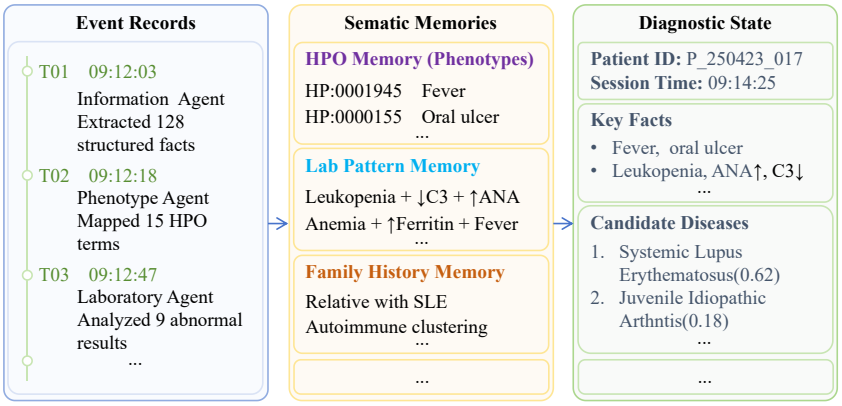

Figure 3: Example patient case showing the structured flow from event records and semantic memories to the diagnostic state and final ranked diagnoses.

Event Records. The event memory records the diagnostic trajectory:

$$\mathcal{E}_t = \{e_1, e_2, \ldots, e_t\} \tag{4}$$

Each event is represented as:

$$e_i = (type_i, agent_i, input_i, output_i, time_i) \tag{5}$$

where $type_i$ indicates the operation type and $output_i$ stores the generated result.

Semantic Memory. Semantic memory stores reusable clinical knowledge:

$$\mathcal{R}_t = \{r_1, r_2, \ldots, r_k\} \tag{6}$$

Each memory item is:

$$r_j = (type_j, content_j, summary_j, score_j, tags_j) \tag{7}$$

where $score_j$ represents confidence and $tags_j$ support retrieval.

Memory Update Mechanism. Given memory $\mathcal{M}_t$ and agent output $\Delta_t^{(i)}$, the middleware updates:

$$\mathcal{M}_{t+1} = (S_{t+1}, \mathcal{E}_{t+1}, \mathcal{R}_{t+1}) \tag{8}$$

where:

$$S_{t+1} = \mathcal{U}_s(S_t, \Delta_t^{(i)}) \tag{9}$$

$$\mathcal{E}_{t+1} = \mathcal{E}_t \cup \{e_t\} \tag{10}$$

$$\mathcal{R}_{t+1} = \mathcal{R}_t \cup \mathcal{U}_r(\Delta_t^{(i)}) \tag{11}$$

where $\mathcal{U}_s$ updates diagnostic states and $\mathcal{U}_r$ extracts reusable knowledge.

Context Construction. Before invoking agent a_i, the middleware constructs an agent-specific context:

$$Z_t^{(i)} = \Phi(S_t, \mathcal{R}_t, a_i) \tag{12}$$

where $\Phi(\cdot)$ retrieves relevant information and constructs the context required by the target agent.

Communication Interface. Agents interact with the middleware through the constructed context:

$$\Delta_t^{(i)} = a_i(Z_t^{(i)}) \tag{13}$$

After receiving $\Delta_t^{(i)}$, the middleware records events, updates diagnostic states, and extracts semantic memories, enabling indirect collaboration through shared memory.

Key Properties. The middleware ensures four properties: **Isolation**, maintaining an independent memory for each patient; **Traceability**, recording complete reasoning trajectories; **Reusability**, storing structured clinical memory; and **Adaptivity**, constructing dynamic agent-specific contexts.

3.2 Agent Modules

Each agent receives the constructed context $Z_t^{(i)}$ and produces structured evidence:

$$\Delta_t^{(i)} = a_i(Z_t^{(i)}) \tag{14}$$

The generated outputs are stored in diagnostic memory for downstream reasoning.

3.3 Agent Details

Information Extraction Agent. It organizes raw records into structured clinical information including demographics, symptoms, laboratory results, medication history, and family history.

Medical Handbook Agent. It retrieves disease knowledge from MSD Medical Manual MSD Manuals (n.d.) and converts retrieved information into structured evidence.

Phenotype Analysis Agent. It extracts phenotypes and maps them to Human Phenotype Ontology (HPO) terms Gargano et al. (2024). SapBERT Liu et al. (2021) improves phenotype normalization.

Family-History Agent. It identifies hereditary clues, affected relatives, consanguinity, and relevant negative findings.

Laboratory Analysis Agent. It interprets laboratory results by identifying abnormalities, missing tests, and clinically meaningful patterns.

3.4 Middleware-Driven Diagnostic Workflow

The diagnostic process iteratively updates patient memory:

$$\mathcal{M}_{t+1} = (S_{t+1}, \mathcal{E}_{t+1}, \mathcal{R}_{t+1}) \tag{15}$$

The controller selects subsequent actions until sufficient evidence is collected.

The final diagnosis is generated as:

$$\hat{Y} = \mathrm{rank}(H_t | S_t, \mathcal{E}_t) \tag{16}$$

where $\hat{Y}$ denotes the ranked diagnosis list.

Algorithm 1 summarizes the complete workflow.

3.5 Advantages of the Middleware Design

The middleware-centered design improves modularity through specialized agents, traceability through event logging, context-aware reasoning through structured memory, and iterative diagnosis through persistent state updates.

4 Experiments

4.1 Datasets

MIMIC-IV-Note Johnson et al. (2023) is derived from real-world clinical notes and contains diagnosis labels with additional clinical fields such as history, laboratory findings, and imaging summaries.

Algorithm 1 Middleware-driven Multi-agent Diagnostic Reasoning

Require: Patient record X, agent set $\mathcal{A}$, diagnostic controller a_d, maximum steps T
Ensure: Ranked diagnosis list $\hat{Y}$

Initialize diagnostic state $S_0 = \{F_0, H_0, U_0, D_0, K_0, E_0\}$
Initialize memory $\mathcal{M}_0 = (S_0, \mathcal{E}_0, \mathcal{R}_0)$
Store raw patient record X as the first memory entry
Set $t \leftarrow 0$ and $ready \leftarrow$ False
while $\neg ready$ and $t < T$ **do**
    Read current diagnostic state S_t from $\mathcal{M}_t$
    Evaluate readiness: $(ready, q_t) \leftarrow a_d(S_t, \mathcal{R}_t)$
    where q_t denotes the diagnostic question or missing evidence identified by the controller.
    if $ready$ **then**
        Produce final ranking: $\hat{Y} \leftarrow \mathrm{rank}(H_t|S_t, \mathcal{E}_t)$

        return $\hat{Y}$
    end if
    Select target agent $a_i \in \mathcal{A}$ based on q_t
    Build agent context: $Z_t^{(i)} \leftarrow \Phi(S_t, \mathcal{R}_t, a_i)$
    Call agent and obtain evidence: $\Delta_t^{(i)} \leftarrow a_i(Z_t^{(i)})$
    Record operation as event e_t
    Update event log: $\mathcal{E}_{t+1} \leftarrow \mathcal{E}_t \cup \{e_t\}$
    Update state: $S_{t+1} \leftarrow \mathcal{U}_s(S_t, \Delta_t^{(i)})$
    Update semantic memory: $\mathcal{R}_{t+1} \leftarrow \mathcal{R}_t \cup \mathcal{U}_r(\Delta_t^{(i)})$
    Refresh memory: $\mathcal{M}_{t+1} \leftarrow (S_{t+1}, \mathcal{E}_{t+1}, \mathcal{R}_{t+1})$
    $t \leftarrow t + 1$
end while
Produce final ranking: $\hat{Y} \leftarrow \mathrm{rank}(H_t|S_t, \mathcal{E}_t)$

return $\hat{Y}$

To create a controlled single-disease diagnosis setting, we filtered cases with only one disease outcome and randomly selected 1,000 cases for evaluation.

RareArena THUMedInfo (2025) focuses on rare disease diagnosis and provides richly annotated cases covering diverse rare disease types. We randomly selected 1,000 cases from its RDC portion to evaluate model performance under challenging low-resource diagnostic conditions.

4.2 MODELS

We evaluated four large language models: DeepSeek-v4 DeepSeek-AI (2024), GPT-5 OpenAI (2023), LLaMA 3.3-70B Meta AI (2024), and Qwen 3-235B Qwen Team (2023). For each model, we compared three diagnostic settings:

1. **Direct diagnosis**: The model predicts diseases directly from patient records without external system support.
2. **Static Workflow**: The model uses all predefined agents with a fixed execution order, without the shared-memory middleware.
3. **DiagMem**: The model uses our middleware-based framework, which dynamically selects agents and maintains patient-specific diagnostic memory.

4.3 EVALUATION METRICS.

We use three ranked retrieval metrics: **R@1**, **R@3**, and **R@5**, indicating whether the correct disease appears in the top-1, top-3, and top-5 predictions, respectively.

Setting	DeepSeek-v4			GPT-5			LLaMA 3.3-70B			Qwen 3-235B		
	R@1	R@3	R@5	R@1	R@3	R@5	R@1	R@3	R@5	R@1	R@3	R@5
Direct	0.689	0.835	0.861	0.780	0.856	0.868	0.643	**0.799**	**0.864**	0.714	0.865	0.890
Static Workflow	0.707	0.829	0.883	0.799	0.864	0.889	0.681	0.791	0.863	0.721	0.848	0.891
DiagMem	**0.726**	**0.847**	**0.901**	**0.805**	**0.870**	**0.908**	**0.698**	0.797	0.828	**0.768**	**0.871**	**0.905**

Table 1: Performance comparison among direct diagnosis, a fixed multi-agent workflow without the shared-memory middleware, and the full DiagMem framework on MIMIC-IV-Note.

Setting	DeepSeek-v4			GPT-5			LLaMA 3.3-70B			Qwen 3-235B		
	R@1	R@3	R@5	R@1	R@3	R@5	R@1	R@3	R@5	R@1	R@3	R@5
Direct	0.434	0.725	0.807	0.692	0.798	0.820	0.365	0.595	0.722	0.413	0.657	0.759
Static Workflow	0.620	0.810	0.843	0.728	0.858	0.894	0.423	0.636	0.731	0.566	0.703	0.761
DiagMem	**0.689**	**0.838**	**0.873**	**0.751**	**0.885**	**0.916**	**0.457**	**0.653**	0.730	**0.587**	**0.735**	**0.788**

Table 2: Performance comparison among direct diagnosis, a fixed multi-agent workflow without the shared-memory middleware, and the full DiagMem framework on RareArena.

4.4 Experimental Results

4.4.1 MIMIC-IV-Note Dataset

Table 1 reports the results on MIMIC-IV-Note. DiagMem improves average R@1 by 4.3 percentage points across four models. The largest gains are observed for LLaMA 3.3-70B and Qwen 3-235B, improving from 0.643 to 0.698 and from 0.714 to 0.768, respectively. GPT-5 and DeepSeek-v4 show smaller improvements, likely because their stronger baselines already capture common clinical patterns from the original records.

Most models also improve in R@3 and R@5. The decrease of LLaMA 3.3-70B in R@5 despite higher R@1 suggests that DiagMem may improve the ranking of the strongest candidate while reshaping lower-ranked alternatives. Overall, memory provides consistent but moderate gains on MIMIC-IV-Note, where backbone models already capture many common patterns.

4.4.2 RareArena Dataset

Table 2 reports results on RareArena. DiagMem improves average R@1 by 14.5 percentage points, substantially exceeding the gains on MIMIC-IV-Note. The largest improvement is achieved by DeepSeek-v4, increasing from 0.434 to 0.689, while Qwen 3-235B improves from 0.413 to 0.587. GPT-5 also benefits, reaching 0.751 R@1 and 0.916 R@5.

The larger gains on RareArena indicate that structured memory is particularly valuable for rare disease diagnosis, where evidence is sparse, distributed, and requires integration across multiple sources.

4.4.3 Comparison with a Static Multi-Agent Workflow

We further compare DiagMem with a static workflow that invokes all agents in a predefined order without maintaining structured patient memory or dynamically selecting the next step. On MIMIC-IV-Note, the static workflow improves average R@1 from 0.707 to 0.727, while DiagMem further raises it to 0.749. On RareArena, the corresponding scores are 0.476, 0.584, and 0.621, respectively.

These results show that specialized agents already provide useful complementary evidence, but the shared-memory middleware yields additional gains. The larger improvement on RareArena suggests that structured state maintenance and adaptive evidence integration are particularly important when diagnostic evidence is sparse and distributed.

Model	Autoimmune & Inflammatory			Genetic & Congenital			Metabolic & Endocrine			Infectious & Neoplastic			Other / Systemic		
	R@1	R@3	R@5	R@1	R@3	R@5	R@1	R@3	R@5	R@1	R@3	R@5	R@1	R@3	R@5
DeepSeek-v4	0.451	0.728	0.835	0.443	0.715	0.772	0.450	0.675	0.725	0.429	0.715	0.801	0.408	0.759	0.843
+ System	0.688	0.826	0.857	0.614	0.754	0.785	0.725	0.775	0.850	0.621	0.743	0.806	0.668	0.806	0.852
GPT-5	0.668	0.787	0.834	0.710	0.813	0.817	0.700	0.800	0.850	0.702	0.785	0.799	0.671	0.810	0.830
+ System	0.723	0.890	0.914	0.758	0.876	0.893	0.700	0.900	0.925	0.747	0.874	0.909	0.789	0.908	0.954
LLaMA 3.3-70B	0.328	0.432	0.453	0.261	0.372	0.388	0.250	0.375	0.475	0.251	0.340	0.355	0.325	0.428	0.446
+ System	0.401	0.685	0.788	0.367	0.571	0.677	0.300	0.550	0.675	0.349	0.557	0.710	0.360	0.589	0.724
Qwen 3-235B	0.464	0.719	0.826	0.390	0.623	0.724	0.375	0.600	0.675	0.423	0.642	0.745	0.372	0.663	0.760
+ System	0.568	0.723	0.809	0.611	0.747	0.783	0.700	0.875	0.900	0.588	0.699	0.755	0.559	0.769	0.800

Table 3: Category-level performance on RareArena across five disease categories.

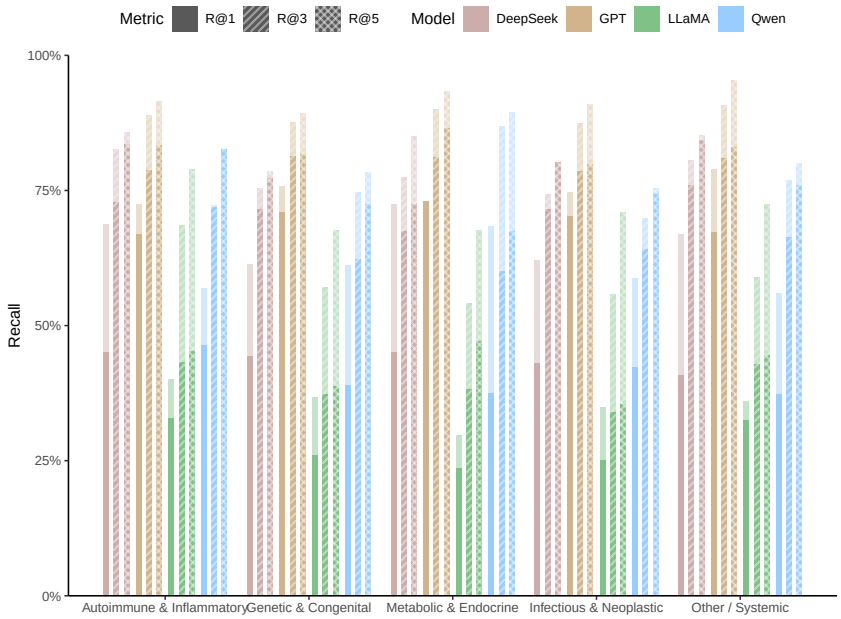

Figure 4: Category-level diagnostic performance (R@1, R@3, R@5) for all four models on the RareArena dataset. The semi-transparent portions allow a clear visual comparison, highlighting the improvement in diagnostic performance achieved by our system.

4.4.4 RareArena Dataset: Multi-Category Evaluation

To analyze performance across different rare disease types, we grouped the RareArena RDC subset into five categories: **Autoimmune & Inflammatory** (224 cases), **Genetic & Congenital** (228 cases), **Metabolic & Endocrine** (40 cases), **Infectious & Neoplastic** (313 cases), and **Other / Systemic** (195 cases).

For each category, we evaluated four models (DeepSeek-v4, GPT-5, LLaMA 3.3-70B, and Qwen 3-235B) using R@1, R@3, and R@5 with and without system augmentation.

Figure 4 visualizes category-level R@1, R@3, and R@5 performance across models.

Overall, system augmentation consistently improves R@1 across most models and categories, with additional gains in R@3 and R@5. These improvements are observed across diverse disease groups, indicating that patient-level memory benefits different types of rare disease reasoning. An exception is GPT-5 on *Metabolic & Endocrine*, where the limited sample size and strong baseline performance leave less room for improvement.

Model	MIMIC			RareArena			Fixed WF
	R@1	R@3	R@5	R@1	R@3	R@5	
DeepSeek	5.43	5.71	5.86	5.79	5.88	6.03	6.00
GPT	**4.27**	**5.37**	**5.73**	**4.28**	**5.46**	**5.94**	6.00
LLaMA	6.20	6.35	6.53	6.27	6.40	7.06	6.00
Qwen	4.63	4.70	5.17	5.45	5.57	5.89	6.00

Table 4: Average number of agent calls for cases correctly diagnosed at different recall levels on MIMIC-IV-Note and RareArena. Fixed WF denotes the fixed workflow, which invokes six agents for every case.

4.5 Agent Usage Across Correctness Levels

We analyze the average number of agent calls at different recall levels. As shown in Table 4, R@1 denotes cases with a correct top prediction, while R@3 and R@5 indicate cases where the correct diagnosis appears within the top-3 or top-5 predictions.

R@1 cases generally require fewer calls than R@3/R@5 cases, suggesting that agent usage reflects case difficulty. Unlike the fixed workflow that always invokes six agents, DiagMem dynamically reduces unnecessary calls when evidence is sufficient and allocates more reasoning to ambiguous cases.

GPT uses the fewest agent calls on average, suggesting stronger initial reasoning or better utilization of information already present in patient records.

4.6 Discussion

DiagMem improves diagnostic ranking on both datasets, with larger gains on RareArena than MIMIC-IV-Note. This indicates that patient-level memory is particularly valuable for rare disease diagnosis, where evidence is sparse and distributed across heterogeneous sources.

The comparison with the static workflow shows that improvements are not solely from using multiple agents. While specialized agents provide complementary evidence, the shared-memory middleware enables additional gains through persistent state maintenance and adaptive evidence integration. On MIMIC-IV-Note, R@1 improves from 0.707 (static workflow) to 0.749 (DiagMem), while on RareArena it increases from 0.584 to 0.621.

Agent usage further shows that DiagMem adjusts reasoning cost according to case difficulty. It avoids unnecessary agent calls for easier cases while allocating additional evidence gathering for ambiguous cases, unlike the fixed workflow that invokes all agents uniformly.

5 Conclusion

We presented *Patient Diagnostic Memory*, a shared-memory framework for multi-agent rare disease diagnosis. Unlike prompt-level diagnostic systems that treat each prediction as an isolated inference, our framework maintains patient-specific diagnostic states, event records, and semantic clinical memories, enabling persistent, traceable, and evidence-grounded reasoning across multiple agents.

Experiments on MIMIC-IV-Note and RareArena with four backbone LLMs demonstrate that DiagMem consistently improves ranked diagnostic accuracy, with particularly larger gains on rare disease cases where evidence is sparse and distributed. The comparison with static multi-agent workflows further shows that the improvements come not only from agent specialization, but also from structured memory maintenance and adaptive evidence integration.

These results highlight the importance of patient-level memory as a foundation for reliable multi-agent medical reasoning. Future work will explore broader clinical settings, prospective validation, and more adaptive memory mechanisms for long-term diagnostic support.

AI USE STATEMENT

In this work, we used generative AI tools for software implementation assistance and editing the manuscript to improve readability. We have not used generative AI tools for developing research methodology, designing experiments, interpreting results, or generating scientific claims. All AI-assisted outputs were carefully reviewed by the authors. We take full responsibility for the final content of this work, including all text, code, claims, and experimental results.

ETHICAL STATEMENT

This work investigates LLM-based diagnostic reasoning for research purposes and is not intended to replace professional medical judgment. The system may produce incorrect or uncertain diagnoses, especially for rare diseases with limited evidence, and its outputs should be reviewed by qualified clinicians.

We do not collect new patient data. MIMIC-IV-Note and RareArena are used under their original de-identification and access requirements. We do not attempt to re-identify individuals or redistribute raw clinical records.

REFERENCES

J. Benito-Lozano, G. Arias-Merino, M. Gómez-Martínez, A. Ancochea-Díaz, A. Aparicio-García, M. Posada de la Paz, and V. Alonso-Ferreira. Diagnostic process in rare diseases: Determinants associated with diagnostic delay. *International Journal of Environmental Research and Public Health*, 19(11):6456, May 2022. doi: 10.3390/ijerph19116456.

Xuanzhong Chen, Xiaohao Mao, Qihan Guo, Lun Wang, Shuyang Zhang, and Ting Chen. Rarebench: Can llms serve as rare diseases specialists? In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4850–4861. ACM, 2024. doi: 10.1145/3637528.3671576. URL https://doi.org/10.1145/3637528.3671576.

DeepSeek-AI. Deepseek-v3 technical report, 2024. URL https://arxiv.org/abs/2412.19437.

Fatoumata Faye, Claudia Crocione, Roberta Anido de Peña, Simona Bellagambi, Luciana Escati Peñaloza, Amy Hunter, Lene Jensen, Cor Oosterwijk, Eva Schoeters, Daniel de Vicente, Laurence Faivre, Michael Wilbur, Yann Le Cam, and Jessie Dubief. Time to diagnosis and determinants of diagnostic delays of people living with a rare disease: results of a rare barometer retrospective patient survey. *European Journal of Human Genetics*, 32:1116–1126, May 2024. doi: 10.1038/s41431-024-01604-z. URL https://doi.org/10.1038/s41431-024-01604-z.

Michael A. Gargano, Nicolas Matentzoglu, Ben Coleman, Eunice B. Addo-Lartey, Anna V. Anagnostopoulos, Joel Anderton, Paul Avillach, Anita M. Bagley, Eduard Bakštein, James P. Balhoff, Gareth Baynam, Susan M. Bello, Michael Berk, Holli Bertram, Somer Bishop, Hannah Blau, David F. Bodenstein, Pablo Botas, Kaan Boztug, Jolana Čady, Tiffany J. Callahan, Rhiannon Cameron, Seth J. Carbon, Francisco Castellanos, J. Harry Caufield, Lauren E. Chan, Christopher G. Chute, Jaime Cruz-Rojo, Noémi Dahan-Oliel, Jon R. Davids, Maud de Dieuleveult, Vinicius de Souza, Bert B. A. de Vries, Esther de Vries, J. Raymond DePaulo, Beata Derfalvi, Ferdinand Dhombres, Claudia Diaz-Byrd, Alexander J. M. Dingemans, Bruno Donadille, Michael Duyzend, Reem Elfeky, Shahim Essaid, Carolina Fabrizzi, Giovanna Fico, Helen V. Firth, Yun Freudenberg-Hua, Janice M. Fullerton, Davera L. Gabriel, Kimberly Gilmour, Jessica Giordano, Fernando S. Goes, Rachel Gore Moses, Ian Green, Matthias Griese, Tudor Groza, Weihong Gu, Julia Guthrie, Benjamin Gyori, Ada Hamosh, Marc Hanauer, Kateřina Hanušová, Yongqun He, Harshad Hegde, Ingo Helbig, Kateřina Holasová, Charles Tapley Hoyt, Shangzhi Huang, Eric Hurwitz, Julius O. B. Jacobsen, Xiaofeng Jiang, Lisa Joseph, Kamyar Keramatian, Bryan King, Katrin Knoflach, David A. Koolen, Megan L. Kraus, Carlo Kroll, Maaike Kusters, Markus S. Ladewig, David Lagorce, Meng Chuan Lai, Pablo Lapunzina, Bryan Laraway, David Lewis-Smith, Xiarong Li, Caterina Lucano, Marzieh Majd, Mary L. Marazita, Victor Martinez-Glez,

Toby H. McHenry, Melvin G. McInnis, Julie A. McMurry, Michaela Mihulová, Caitlin E. Millett, Philip B. Mitchell, Veronika Moslerová, Kenji Narutomi, Shahrzad Nematollahi, Julian Nevado, Andrew A. Nierenberg, Nikola Novák Čajbiková, John I. Nurnberger, Soichi Ogishima, Daniel Olson, Abigail Ortiz, Harry Pachajoa, Guiomar Perez de Nanclares, Amy Peters, Tim Putman, Christina K. Rapp, Ana Rath, Justin Reese, Lauren Rekerle, Angharad M. Roberts, Suzy Roy, Stephan J. Sanders, Catharina Schuetz, Eva C. Schulte, Thomas G. Schulze, Martin Schwarz, Katie Scott, Dominik Seelow, Berthold Seitz, Yiping Shen, Morgan N. Similuk, Eric S. Simon, Balwinder Singh, Damian Smedley, Cynthia L. Smith, Jake T. Smolinsky, Sarah Sperry, Elizabeth Stafford, Ray Stefancsik, Robin Steinhaus, Rebecca Strawbridge, Jagadish Chandrabose Sundaramurthi, Polina Talapova, Jair A. Tenorio Castano, Pavel Tesner, Rhys H. Thomas, Audrey Thurm, Marek Turnovec, Marielle E. van Gijn, Nicole A. Vasilevsky, Markéta Vlčková, Anita Walden, Kai Wang, Ron Wapner, James S. Ware, Addo A. Wiafe, Samuel A. Wiafe, Lisa D. Wiggins, Andrew E. Williams, Chen Wu, Margot J. Wyrwoll, Hui Xiong, Nefize Yalin, Yasunori Yamamoto, Lakshmi N. Yatham, Anastasia K. Yocum, Allan H. Young, Zafer Yüksel, Peter P. Zandi, Andreas Zankl, Ignacio Zarante, Miroslav Zvolský, Sabrina Toro, Leigh C. Carmody, Nomi L. Harris, Monica C. Munoz-Torres, Daniel Danis, Christopher J. Mungall, Sebastian Köhler, Melissa A. Haendel, and Peter N. Robinson. The human phenotype ontology in 2024: phenotypes around the world. *Nucleic Acids Research*, 52(D1):D1333–D1346, January 2024. doi: 10.1093/nar/gkad1005. URL https://doi.org/10.1093/nar/gkad1005.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023. doi: 10.1038/s41597-022-01899-x. URL https://doi.org/10.1038/s41597-022-01899-x.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *EMNLP*, 2020. doi: 10.48550/arXiv.2004.04906. URL https://doi.org/10.48550/arXiv.2004.04906. arXiv preprint arXiv:2004.04906.

S. Köhler, M. H. Schulz, P. Krawitz, S. Bauer, S. Dölken, C. E. Ott, C. Mundlos, D. Horn, S. Mundlos, and P. N. Robinson. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *American Journal of Human Genetics*, 85(4):457–464, October 2009. doi: 10.1016/j.ajhg.2009.09.003.

Wuyang Lan, Wenzheng Wang, Changwei Ji, Guoxing Yang, Yongbo Zhang, Xiaohong Liu, Song Wu, and Guangyu Wang. Clinicalgpt-r1: Pushing reasoning capability of generalist disease diagnosis with large language model. *arXiv preprint arXiv:2504.09421*, 2025. doi: 10.48550/arXiv.2504.09421. URL https://doi.org/10.48550/arXiv.2504.09421. 8 pages, 6 figures.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems 33*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large language model society. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 51991–52008. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a3621ee907def47c1b952ade25c67698-Paper-Conference.pdf.

Wenliang Li, Rui Yan, Xu Zhang, Li Chen, Hongji Zhu, Jing Zhao, Junjun Li, Mengru Li, Wei Cao, Zihang Jiang, Wei Wei, Kun Zhang, and Shaohua Kevin Zhou. MACD: Multi-agent clinical diagnosis with self-learned knowledge for LLM. *arXiv preprint arXiv:2509.20067*, 2025. doi: 10.48550/arXiv.2509.20067. URL https://arxiv.org/abs/2509.20067.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-alignment pretraining for biomedical entity representations. In *NAACL 2021, Camera-Ready Version*, 2021. doi: 10.48550/arXiv.2010.11784. URL https://doi.org/10.48550/arXiv.2010.11784. arXiv preprint arXiv:2010.11784.

Yuxing Lu, Xukai Zhao, and Jinzhuo Wang. ClinicalRAG: Enhancing clinical decision support through heterogeneous knowledge retrieval. In Sha Li, Manling Li, Michael JQ Zhang, Eunsol Choi, Mor Geva, Peter Hase, and Heng Ji (eds.), *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pp. 64–68, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.knowllm-1.6. URL https://aclanthology.org/2024.knowllm-1.6/.

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, Le Hou, Yong Cheng, Yun Liu, S. Sara Mahdavi, Sushant Prakash, Anupam Pathak, Christopher Semturs, Shwetak Patel, Dale R. Webster, Ewa Dominowska, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, Jake Sunshine, Alan Karthikesalingam, and Vivek Natarajan. Towards accurate differential diagnosis with large language models. *Nature*, 642:451–457, April 2025. doi: 10.1038/s41586-025-08869-4. URL https://www.nature.com/articles/s41586-025-08869-4.

Meta AI. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. doi: 10.48550/arXiv.2407.21783. URL https://arxiv.org/abs/2407.21783.

MSD Manuals. MSD Manual Professional Version. https://www.msdmanuals.com/professional, n.d. Accessed: 2026-04-30.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. doi: 10.48550/arXiv.2303.08774. URL https://arxiv.org/abs/2303.08774.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023. doi: 10.48550/arXiv.2310.08560. URL https://arxiv.org/abs/2310.08560. Code and data available.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, pp. 1–22, San Francisco, CA, USA, 2023. ACM. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Qwen Team. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. doi: 10.48550/arXiv.2309.16609. URL https://arxiv.org/abs/2309.16609.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomašev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle K. Barral, Dale R. Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31:943–950, January 2025. doi: 10.1038/s41591-024-03423-7. URL https://www.nature.com/articles/s41591-024-03423-7.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 599–621, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.33. URL https://aclanthology.org/2024.findings-acl.33/.

THUMedInfo. Rarearena: A comprehensive rare disease diagnostic dataset with nearly 50,000 patients covering more than 4,000 diseases. Hugging Face Datasets, 2025. Version: main (commit 85e5d3a), accessed 2025-12-30.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, Yueming Jin, and Vicente Grau. Medical graph rag: Evidence-based medical large language model via graph retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28443–28467, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1381. URL https://aclanthology.org/2025.acl-long.1381/.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

W. Zhao, C. Wu, Y. Fan, and et al. An agentic system for rare disease diagnosis with traceable reasoning. *Nature*, 651:775–784, 2026. doi: 10.1038/s41586-025-10097-9. URL https://doi.org/10.1038/s41586-025-10097-9.

Qiaoyu Zheng, Yuze Sun, Chaoyi Wu, Weike Zhao, Pengcheng Qiu, Yongguo Yu, Kun Sun, Yanfeng Wang, Ya Zhang, and Weidi Xie. End-to-end agentic rag system training for traceable diagnostic reasoning. *arXiv preprint arXiv:2508.15746*, 2025. doi: 10.48550/arXiv.2508.15746. URL https://arxiv.org/abs/2508.15746.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. *arXiv preprint arXiv:2305.10250*, 2023. doi: 10.48550/arXiv.2305.10250. URL https://arxiv.org/abs/2305.10250. 10 pages.

Jiayuan Zhu, Jiazhen Pan, Yuyuan Liu, Fenglin Liu, and Junde Wu. Ask patients with patience: Enabling LLMs for human-centric medical dialogue with grounded reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2846–2857, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.142. URL https://aclanthology.org/2025.emnlp-main.142/.