Rare Disease Diagnosis Agent with Decoupled Workflows and Knowledge-Driven Self-Evaluation

Yunlu Yan[1*] **Yawen Huang**[2*] **Xian Wu**[2†] **Lei Zhu**[1,3†]
[1] The Hong Kong University of Science and Technology (Guangzhou)
[2] Tencent Jarvis Lab
[3] The Hong Kong University of Science and Technology

Abstract

Rare disease diagnosis is a fundamental challenge due to heterogeneous and overlapping clinical phenotypes. While recent LLM-based agentic systems have demonstrated promise by integrating external medical tools and knowledge, they typically rely on large-scale or commercial LLMs, limiting their practical deployment in resource-constrained settings. In this work, we observe that their performance significantly degrades when using small-scale LLMs, due to entangled workflows and unreliable multi-evidence aggregation. To address this issue, we propose **RADAR**, a rare disease diagnostic framework designed for small-scale LLMs. RADAR adopts a divide-and-conquer paradigm that decouples heterogeneous diagnostic evidence into specialized workflows, reducing reasoning complexity and improving robustness. It further introduces a knowledge-driven self-evaluation mechanism that performs evidence-aware reasoning over structured disease knowledge to produce interpretable reliability scores by identifying key supporting phenotypes and conflicting evidence. Finally, a multi-evidence fusion strategy integrates outputs from multiple sources based on reliability and clinical agreement, resolving conflicts and producing stable diagnostic rankings. Experiments on three benchmarks show that RADAR consistently outperforms various state-of-the-art baselines, including general and medical LLMs and agentic systems.

1 Introduction

Rare diseases [33] are individually rare (affecting fewer than 1 in 2,000 individuals), yet collectively encompass more than 7,000 distinct conditions, with new diseases continuously being identified each year, impacting over 300 million people worldwide [15]. Patients with rare diseases often exhibit complex, heterogeneous, and multisystem phenotypes that require multidisciplinary expertise for diagnosis, leading to prolonged diagnostic timelines and potential delays in optimal treatment [26]. Developing artificial intelligence (AI) systems for rare disease diagnosis to improve diagnostic accuracy and efficiency has emerged as a promising direction. However, traditional deep learning methods [28, 42, 18] are constrained by the scarcity of rare disease cases, the continuously expanding disease spectrum, and the inherent complexity of rare disease diagnosis, resulting in limited effectiveness in real-world clinical settings.

Recent agentic systems have demonstrated strong performance in healthcare [16, 32, 43, 35]. Unlike traditional end-to-end paradigms, these methods leverage large language models (LLMs) to integrate multiple speicalized diagnostic tools and external APIs, allowing access to diverse and up-to-date medical knowledge. This brings model reasoning closer to real-world clinical decision-making

*Yunlu Yan and Yawen Huang contributed equally to this work. This work is done during Yunlu Yan's internship at Tencent Jarvis Lab.
†Corresponding authors: Xian Wu and Lei Zhu.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

workflows. Motivated by this, DeepRare [48], a pioneering work for rare disease diagnosis, leverages LLMs to orchestrate multiple diagnostic tools (*e.g.*, PhenoBrain [23] and PubCaseFinder [10]), case databases, external knowledge APIs (*e.g.*, Google and Bing), and rare disease knowledge bases (*e.g.*, Orphanet [36] and OMIM [2]) within a unified framework, significantly improving diagnostic accuracy while enabling traceable decision-making pathways. Despite the advances, similar to other medical agentic systems, DeepRare relies on either closed-source commercial LLMs (*e.g.*, GPT-4o [1]) or open-source large-scale LLMs (*e.g.*, DeepSeek-V3 [21]). The former introduces privacy concerns regarding the transmission of sensitive patient data to external servers, while the latter imposes significant barriers to real-world deployment due to prohibitive computational and memory requirements (*e.g.*, DeepSeek-V3 contains 671B parameters). This creates a critical deployment gap: *while developing nations often face a higher burden of rare disease cases, hospitals in these regions frequently lack the economic resources to support such large-scale infrastructure*.

A potential solution to the above challenges is to adopt small-scale LLMs. However, we observe that directly replacing the backbone of DeepRare with a small-scale LLM leads to a significant performance degradation. Through in-depth analysis, we identify two key limitations of the DeepRare agentic pipeline under this setting. ***First***, DeepRare aggregates outputs from multiple diagnostic tools, case-based evidence, and external API-retrieved knowledge into a unified context to produce the final diagnostic result (see Figure 1 (a)). This design implicitly requires the underlying LLM to simultaneously perform multi-source evidence aggregation, conflict resolution, and complex reasoning over heterogeneous information. Such a monolithic pipeline entangles multiple reasoning tasks within a single model, imposing a substantial cognitive burden. While large-scale LLMs can partially handle this complexity, the design becomes particularly ineffective for small-scale LLMs due to their limited reasoning capacity and long-context understanding ability (see Figure 1 (b)). ***Second***, during the reflection stage, DeepRare inputs candidate disease knowledge from rare disease databases together with patient phenotypes into an LLM for binary (True/False) judgment. This formulation is highly susceptible to confirmation bias, encouraging shallow phenotype matching and causing the model to prematurely reject correct diagnoses due to partial mismatches (see Figure 1 (c)). Moreover, such a paradigm is fundamentally misaligned with real-world clinical reasoning, where diagnoses are typically supported by a subset of key phenotypes and refined through the identification of conflicting evidence, rather than strict full-pattern matching.

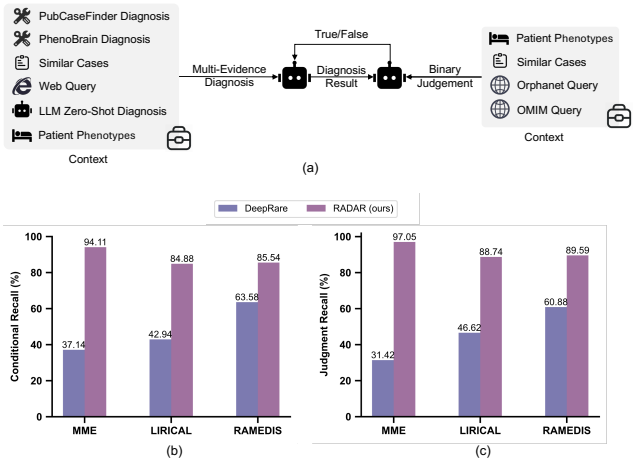


Figure 1: **Overview** [48] of DeepRare with phenotypes-only input. (b) **Conditional Recall (%)** comparison between DeepRare and ours: proportion of cases in which the final diagnosis is correct (Recall ≤ 5), among those where at least one diagnostic tool successfully predicts the ground-truth disease within top-5. This suggests that DeepRare is limited in effectively aggregating evidence and resolving conflicts among different diagnostic tools. (c) **Judgment Recall (%)** measures whether the judgment mechanism in DeepRare and our method correctly identifies the ground-truth disease during the reasoning process. This further reveals that DeepRare's judgment mechanism is unreliable in evaluating candidate diagnoses, resulting in rejection of correct predictions.

In this work, we propose RADAR, a rare disease diagnosis agent with small-scale LLMs. RADAR adopts a divide-and-conquer paradigm, decomposing heterogeneous multi-source diagnosis into dedicated workflows to reduce reasoning complexity and improve decision reliability. Furthermore, we propose a knowledge-driven self-evaluation strategy to evaluate the reliability of different diagnostic sources. It retrieves structured disease knowledge from rare disease databases and performs evidence-aware and graded reasoning to identify key supporting features while detecting conflicting evidence, producing a reliability score together with interpretable decision rationales. Based on these structured evaluations, we further introduce a multi-evidence fusion strategy to integrate heterogeneous

diagnostic outputs from different sources. The fusion process jointly considers reliability scores, cross-source agreement, and clinical rationales to resolve conflicts and produce a consistent ranking of candidate diseases. The final aggregated results provide interpretable and clinically grounded diagnostic decisions.

In a nutshell, our contributions are summarized as follows:

- We identify the limitations of existing agentic diagnostic pipelines under small-scale LLM settings, highlighting the issues of workflow entanglement and confirmation bias that hinder reliable multi-step clinical reasoning.
- We propose RADAR, a small-scale LLM-based rare disease diagnosis framework that decouples heterogeneous diagnosis into specialized workflows, introduces a knowledge-driven self-evaluation strategy for reliability scoring with interpretable rationales, and integrates multi-source outputs through an evidence-aware fusion process.
- We conduct experiments on three rare disease diagnosis benchmarks, demonstrating that RADAR significantly outperforms various state-of-the-art general and medical LLMs and agentic systems. In addition, under the same small-scale underlying LLM, RADAR consistently achieves substantial improvements over DeepRare in both recall and efficiency.

2 Related Work

2.1 AI for Rare Disease Diagnosis

AI-assisted diagnosis can significantly improve diagnostic accuracy and efficiency for rare diseases, thereby reducing diagnostic delays and improving patient outcomes. Early studies [17, 31, 10, 44, 46, 25] primarily relied on human phenotype ontology (HPO)-based methods to standardize clinical descriptions and employed semantic similarity measures to rank candidate diseases according to phenotypic overlap. With the advent of deep learning, natural language processing (NLP)-based methods such as Clinphen [7] and PhenoBrain [23] have significantly enhanced the automation of phenotype extraction and differential diagnosis from electronic health records. However, these methods often operate in a static manner and struggle with the ambiguity inherent in real-world clinical applications. Recently, Chen *et al.* [4] proposed RareBench, introducing a dynamic few-shot prompting method to improve LLM-based diagnostic performance. Despite its effectiveness, this approach primarily relies on the model's internal parametric knowledge and lacks the capability to interact with specialized diagnostic tools or live medical knowledge bases, thereby limiting its ability to resolve complex diagnostic dilemmas that require multi-source evidence fusion.

2.2 Medical Agentic System

The recent rise of LLMs has facilitated a shift from static paradigms to agentic decision-making, demonstrating strong performance in various applications [20]. By incorporating planning [45, 38], memory [49, 40], self-evaluation [39, 29, 22] and tool-using [37, 12] capabilities, LLM-based agents are able to perform complex tasks through multi-step reasoning and interaction with external environments. Motivated by this, medical agentic systems [34, 47, 5, 50] leverage LLMs as central orchestrators to interact with external tools, clinical APIs, and biomedical knowledge bases, enabling multi-step reasoning over heterogeneous clinical information [16, 32, 43, 35, 19]. Compared to traditional end-to-end diagnostic models, these systems [9, 35] more closely resemble real-world clinical workflows by iteratively gathering evidence and refining diagnostic hypotheses. In rare disease diagnosis, recent work such as DeepRare [48] further extends this paradigm by integrating multiple diagnostic tools, case databases, and medical knowledge sources within a unified agentic framework. Despite achieving promising performance, existing medical agentic systems are typically built upon large-scale or commercial LLMs, which raises concerns regarding privacy, computational cost, and deployment feasibility in real-world clinical settings. In contrast, our work aims to address these limitations by introducing a workflow-decoupled agentic framework with a knowledge-driven self-evaluation strategy, enabling reliable and efficient rare disease diagnosis with small-scale LLMs.

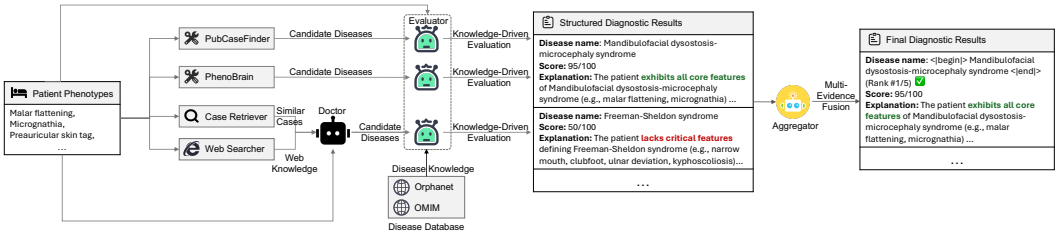


Figure 2: **Illustration of RADAR**, decoupling the diagnostic process into independent workflows. The ***Doctor Agent*** conducts initial diagnosis by integrating patient information with external medical knowledge, emulating human-like clinical reasoning. The ***Evaluator Agent*** applies a knowledge-driven evaluation strategy to assess heterogeneous diagnostic outputs and generate structured evaluation signals. The ***Aggregator Agent*** performs multi-evidence fusion and reasoning over both diagnostic outputs and evaluation signals, and derives the final diagnostic result.

3 Methodology

3.1 Problem Statement

In this work, we follow the phenotypes-only setting of DeepRare, which focuses on phenotype-based differential diagnosis. The goal of differential diagnosis for rare diseases is to accurately identify a target rare disease by distinguishing it from other disorders with overlapping clinical manifestations. Given a patient's clinical phenotype information, represented as an HPO sequence $\boldsymbol{Q} = \{\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_m\}$, we aim to predict a ranked list of candidate rare diseases $\boldsymbol{\mathcal{A}} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n\}$, where higher-ranked candidates are more likely to be the correct diagnosis. The ground-truth disease is expected to be ranked as high as possible, thereby facilitating a more efficient diagnostic process.

3.2 Decoupled Diagnostic Workflows

Figure 2 illustrates the overall architecture of RADAR. Our framework integrates both conventional diagnostic tools and retrieval-augmented generation (RAG)-based reasoning within a unified agentic system. Unlike prior work [48], which concatenates heterogeneous diagnostic evidence into a single context for LLM inference, RADAR decouples the diagnostic process into a set of independent yet collaborative workflows. This decoupled design mitigates potential conflicts among heterogeneous diagnostic signals and reduces the complexity of long-context reasoning, thereby better aligning with the practical constraints of small-scale LLMs, particularly their limited long-context modeling capacity and reasoning robustness.

Specifically, for a given patient phenotype, RADAR employs two complementary diagnostic workflows in parallel, *i.e.*, diagnostic tools and a RAG-based agent, to make the diagnosis from different perspectives. This multi-view design enables the model to capture diverse diagnostic evidence from different sources, thereby improving robustness and reducing the risk of single-source bias.

Diagnostic Tools. We employ two traditional HPO-based methods, **PubCaseFinder**[3] and **PhenoBrain**[4], as external diagnostic engines. These tools function as web-based services that accept patient phenotypes, represented by HPO terms, and return a ranked list of rare diseases. We select the top-n diseases as candidate diseases. Let $\boldsymbol{f}_{\mathrm{PCF}}(\cdot)$ and $\boldsymbol{f}_{\mathrm{PB}}(\cdot)$ denote the mapping functions of PubCaseFinder and PhenoBrain, respectively. The diagnostic process is expressed as:

$$\boldsymbol{\mathcal{A}}_{\mathrm{PCF}} = \boldsymbol{f}_{\mathrm{PCF}}(\boldsymbol{Q}), \quad \boldsymbol{\mathcal{A}}_{\mathrm{PB}} = \boldsymbol{f}_{\mathrm{PB}}(\boldsymbol{Q}). \tag{1}$$

RAG-based Agent. The ***Doctor Agent*** is equipped with two retrieval tools, *i.e.*, **Web Searcher** and **Case Retriever**. The web searcher retrieves relevant medical knowledge from the Bing[5] search engine based on patient phenotypes. The retrieved content serves as auxiliary evidence, grounding and enhancing downstream diagnostic reasoning. The case retriever leverages the case database

[3] https://pubcasefinder.dbcls.jp/api
[4] http://www.phenobrain.cs.tsinghua.edu.cn
[5] https://www.bing.com/

provided by DeepRare and adopts MedCPT [14] as a similarity-based retrieval model to identify patient cases with phenotypes similar to the query patient. These retrieved cases provide case-based evidence, enabling more informed and context-aware reasoning. These evidences are integrated to support the RAG-based diagnostic process (see ***Prompt 1*** in Appendix A). This design closely aligns with real-world clinical practice, where physicians rely on both accumulated clinical experience (cases) and domain knowledge (medical literature) to make diagnostic decisions. Let $\mathcal{R}_{\text{case}}(\cdot)$ and $\mathcal{R}_{\text{web}}(\cdot)$ denote the case retriever and web searcher, respectively, and let $\boldsymbol{f}_{\text{Doc}}(\cdot)$ denote the Doctor Agent. The diagnostic process is formulated as:

$$\mathcal{A}_{\text{Doc}} = \boldsymbol{f}_{\text{Doc}}(\boldsymbol{Q}; \mathcal{T}; \mathcal{M}), \quad \mathcal{M} = \mathcal{R}_{\text{case}}(\boldsymbol{Q}), \quad \mathcal{T} = \mathcal{R}_{\text{web}}(\boldsymbol{Q}), \tag{2}$$

where $\mathcal{M}$ denotes retrieved similar cases and $\mathcal{T}$ denotes retrieved web-based medical knowledge.

3.3 Knowledge-Driven Self-Evaluation

The self-evaluation mechanism evaluates the reliability and consistency of diagnostic outputs by grounding them in external medical evidence and facilitates subsequent evidence fusion through selection and ranking to resolve conflicts among heterogeneous diagnostic results. It is implemented by the ***Evaluator Agent***, which is equipped with retrieval tools to access structured disease knowledge from rare disease databases, including Orphanet[6] and OMIM[7], via HTTP-based queries. Figure 3 (b) shows an example of the queried knowledge from rare disease databases, including the disease name, definition, and associated phenotypes. These phenotypes appear with different frequencies (see Figure 3 (a)), reflecting their varying importance for diagnosis. Prior work formulates diagnostic verification as a binary (True/False) prediction. This formulation is prone to confirmation bias, as it relies on strict phenotype matching and may incorrectly reject valid diagnoses due to partial mismatches. Moreover, it is misaligned with real clinical reasoning, where diagnosis depends on key supporting phenotypes and the analysis of conflicting evidence.

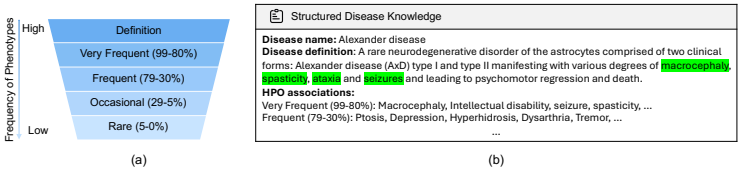


Figure 3: **Illustration** of queried knowledge from rare disease databases. (a) Types of phenotype associations for rare diseases. (b) An example of structured disease knowledge, where marked phenotypes denote the defining features of the disease.

To this end, we propose a knowledge-driven self-evaluation method that models diagnostic judgment as an evidence-aware and graded reasoning by a chain-of-thought prompt (see ***Prompt 2*** in Appendix A). Specifically, instead of treating all phenotypes equally, we assign different roles to phenotypes based on their occurrence frequency. Definition and highly frequent phenotypes serve as core indicators of the disease and contribute to the overall reliability in a soft manner, where missing features lead to proportional penalties rather than hard exclusion. Moderately frequent phenotypes provide complementary evidence to refine the diagnostic profile, while less frequent phenotypes act as supportive cues that contribute positively when present but are not penalised when absent. In addition, we explicitly model conflict awareness by identifying phenotypes that are logically or clinically incompatible with the disease (*e.g.*, hyperglycemia *vs.* hypoglycemia). Such conflicts are treated as strong negative signals and incur larger penalties than missing features, reflecting their higher diagnostic significance. Based on these principles, the Evaluator Agent performs a structured assessment by jointly considering (1) the presence of key defining features from the disease definition, (2) the weighted coverage of phenotypes across frequency tiers, (3) the overall consistency of the clinical pattern, and (4) potential conflicts. The final output (see Figure 2) is a continuous reliability score accompanied by concise reasoning, which serves as structured evidence for downstream aggregation and decision-making. Let $\mathcal{R}_{\text{rare}}(\cdot)$ and $\boldsymbol{f}_{\text{eval}}(\cdot)$ denote the mapping functions of the rare disease knowledge retriever and the Evaluator Agent, respectively. The evaluation process can be formulated as:

$$\hat{\mathcal{A}}_{\text{eval}} = \{(\boldsymbol{a}_i, \boldsymbol{s}_i, \boldsymbol{e}_i)\}_{i=1}^{3n}, \quad (\boldsymbol{s}_i, \boldsymbol{e}_i) = \boldsymbol{f}_{\text{eval}}(\boldsymbol{a}_i; \mathcal{R}_{\text{rare}}(\boldsymbol{a}_i)), \quad \boldsymbol{a}_i \in \mathcal{A}_{\text{PCF}} \cup \mathcal{A}_{\text{PB}} \cup \mathcal{A}_{\text{Doc}}, \tag{3}$$

[6]https://www.orpha.net/
[7]https://www.omim.org/

where s_i denotes the judgment score and e_i represents the corresponding explanation.

3.4 Multi-Evidence Fusion

Given the heterogeneous diagnostic outputs from multiple sources, including traditional diagnostic tools and the RAG-based agent, we perform multi-evidence fusion to derive the final diagnostic decision. Each source produces a set of candidate diseases with an associated reliability score and corresponding judgment explanation. To integrate these heterogeneous outputs, we first collect all candidate diseases from different sources and identify the set of unique diseases. For each disease, we aggregate two types of information: (1) its reliability score produced by the Evaluator Agent, and (2) its support count, defined as the number of sources that predict the disease. We then perform ranking based on a hierarchical decision strategy. Specifically, diseases are primarily ranked according to their reliability scores. When multiple diseases share similar scores, we further prioritize those supported by a larger number of sources, reflecting higher cross-source consistency. In cases where both the scores and support counts are comparable, we leverage the associated judgment explanation to perform a fine-grained comparison based on how well the disease is supported by the patient phenotypes. This process enables effective integration of complementary evidence from different diagnostic sources and balances model confidence with cross-source agreement. The final output is a top-n ranked list of candidate diseases, together with their reliability scores and corresponding explanations, providing interpretable diagnostic results. These steps are unified under a chain-of-thought reasoning process (see ***Prompt 3*** in Appendix A), allowing the ***Aggregator Agent*** to reason over multi-source evidence and produce the final decision. Let the Aggregator Agent be denoted as $\boldsymbol{f}_{\text{agg}}(\cdot)$. The aggregation process is formulated as:

$$\hat{\boldsymbol{\mathcal{A}}}_{\text{final}} = \boldsymbol{f}_{\text{agg}}(\hat{\boldsymbol{\mathcal{A}}}_{\text{eval}}). \tag{4}$$

4 Experiments

4.1 Experimental Setup

Datasets. We use threes datasets, MME (40 cases), LIRICAL (370 cases) and RAMEDIS (624 cases) from RareBench [4]. These datasets are constructed from clinical literature and case reports and further manually validated. They provide patient phenotypes in the form of HPO terms and reflect realistic open-ended diagnostic scenarios. We adopt GPT-5.1 [30] as an automatic evaluator to assess the correctness of predicted diagnoses (see ***Prompt 4*** in Appendix A). Specifically, a prediction is considered correct if the ground-truth disease appears within the top-ranked results. We report recall-1 (R@1), recall-3 (R@3), and recall-5 (R@5) to measure diagnostic performance.

Baselines. We compare RADAR against various baselines, categorized into four groups: ❶ **General LLMs**: We compare with six state-of-the-art commercial LLMs including DeepSeek-R1 [13], DeepSeek-V3 [21], Gemini-2.5-pro [6], Claude-sonnet-4 [3], GPT-4-turbo [1] and GPT-5.1 [30]. ❷ **General agentic systems**: Our comparison includes two prominent commercial agentic frameworks, *i.e.*, GLM-5 [11] and Kimi-K2.6 [24]. ❸: **Medical LLMs**: We incorporate two representative open-source medical models, *i.e.*, MedGemma-27B-Text [27] and Baichuan-M2-32B [8]. ❹: **Medical agentic systems**: We further compare with specialized agent-based methods, *i.e.*, MDAgent [16] and DeepRare [48].

Implementation Details. Our experiment is conducted using a single NVIDIA A6000 GPU. For a fair comparison, we adopt the same underlying LLM (Qwen3-8B [41]) for MDAgent, DeepRare, and our method.

4.2 Main Results

Comparison with General LLMs and Agentic Systems. We first compare RADAR with general LLMs and agentic systems on MME, LIRICAL, and RAMEDIS. As shown in Table 1, recent commercial LLMs such as GPT-4-Turbo, GPT-5.1, Gemini-2.5-Pro, and Claude-Sonnet-4 achieve relatively strong performance, indicating their general reasoning capability. However, their performance remains limited on rare disease diagnosis, especially on MME, where even the best LLM (Gemini-2.5-Pro) only achieves 37.50% R@1. This is likely because rare disease data are scarce in their pre-trained data, limiting the models' ability to capture long-tail medical knowledge. General

Table 1: **Comparison** with commercial LLMs and medical agentic systems on **MME**, **LIRICAL** and **RAMEDIS**. Best results are marked in **bold**.

Method	MME			LIRICAL			RAMEDIS		
	R@1	R@3	R@5	R@1	R@3	R@5	R@1	R@3	R@5
General LLMs									
Qwen3-8B	0.00	10.00	17.50	16.48	31.89	40.27	29.80	57.37	63.14
DeepSeek-R1	25.00	37.50	42.50	35.40	48.10	53.51	29.64	54.16	63.94
DeepSeek-V3	5.00	20.00	27.50	33.51	49.18	53.51	35.41	59.77	65.38
Gemini-2.5-pro	37.50	45.00	52.50	35.94	48.64	53.51	35.73	56.25	65.06
Claude-sonnet-4	2.50	22.50	30.00	31.35	43.78	49.45	36.85	66.98	77.56
GPT-4-turbo	10.00	20.00	45.00	35.13	46.48	53.78	36.53	61.21	70.51
GPT-5.1	17.50	35.00	50.00	24.59	40.00	48.10	28.52	49.67	65.06
General Agentic Systems									
GLM-5	42.50	55.00	60.00	28.37	45.40	51.08	29.48	51.28	61.21
Kimi-K2.6	35.00	50.00	55.00	30.81	43.24	49.18	36.37	59.61	67.94
Medical LLMs									
MedGemma-27B-Text	0.00	7.50	12.50	24.05	36.75	41.89	40.06	62.01	72.27
Baichuan-M2-32B	2.50	7.50	7.50	24.59	38.10	42.43	33.33	53.84	64.42
Medical Agentic Systems									
MDAgent (Qwen3-8B)	5.00	22.50	25.00	11.89	23.51	31.89	14.42	43.91	56.57
DeepRare (Qwen3-8B)	17.50	30.00	37.50	16.75	28.91	39.72	24.03	44.39	57.53
RADAR (Qwen3-8B)	**57.50**	**72.50**	**82.50**	**42.97**	**63.24**	**74.05**	**41.66**	**68.26**	**78.52**

agentic systems, including GLM-5 and Kimi-K2.6, further improve the preformance by incorporating tool-use or multi-step reasoning. Nevertheless, their gains are still insufficient to address the complexity of rare disease diagnosis, suggesting that general agentic systems lack domain-specific medical grounding. In contrast, RADAR significantly outperforms all general LLMs and agentic systems across all datasets. For instance, it improves R@1 from 37.50% to **57.50**% on MME, from 35.94% to **42.97**% on LIRICAL, and from 36.85% to **41.66**% on RAMEDIS. These improvements confirm that RADAR's approach of incorporating multi-source evidence and expert medical knowledge is essential for accurate rare disease diagnosis.

Comparison with Medical LLMs and Agentic Systems. We further compare RADAR with medical LLMs and agent-based diagnostic systems. Medical LLMs such as MedGemma-27B-Text and Baichuan-M2-32B are obtained by fine-tuning general LLMs on medical data. Consequently, their diagnostic capabilities remain largely reliant on parametric knowledge implicitly encoded during training. Due to the long-tail distribution and inherent complexity of rare diseases, they struggle to generalize effectively across diverse clinical scenarios. For example, Baichuan-M2-32B achieves 64.42% R@5 on RAMEDIS, but only 7.50% R@5 on MME, indicating a significant performance drop under more challenging and rare diseases. For medical agentic systems, DeepRare achieves stronger performance than MDAgent, benefiting from specialized diagnostic tools and rare disease knowledge. However, these methods often rely on powerful large-scale LLMs as their underlying backbone, making them less effective with small-scale LLMs, as they struggle to handle the tightly coupled diagnostic process. In contrast, the consistent improvements of RADAR over both medical LLMs and medical agentic baselines demonstrate its effectiveness across diverse rare disease scenarios. By decoupling diagnosis into complementary workflows and aggregating heterogeneous diagnostic outputs, RADAR effectively mitigates conflicts among different sources of evidence and improves robustness in rare disease diagnosis.

4.3 Ablation Studies

Comparison with Single-evidence Diagnosis. To evaluate the effectiveness of the multi-evidence diagnostic paradigm in RADAR, we analyze the performance of its three internal single-evidence diagnosis pathways, *i.e.*, PubCaseFinder, PhenoBrain, and Doctor Agent, as shown in Figure 4. We

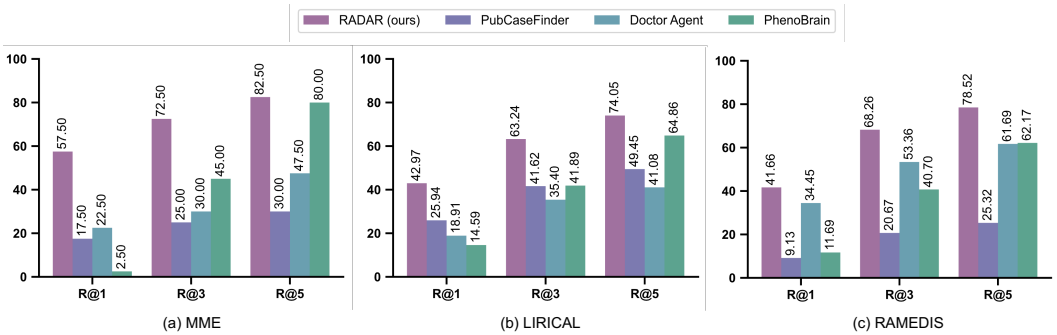


Figure 4: **Comparison** against single-evidence diagnosis methods, *i.e.*, PubCaseFinder, PhenoBrain and Doctor Agent on three datasets.

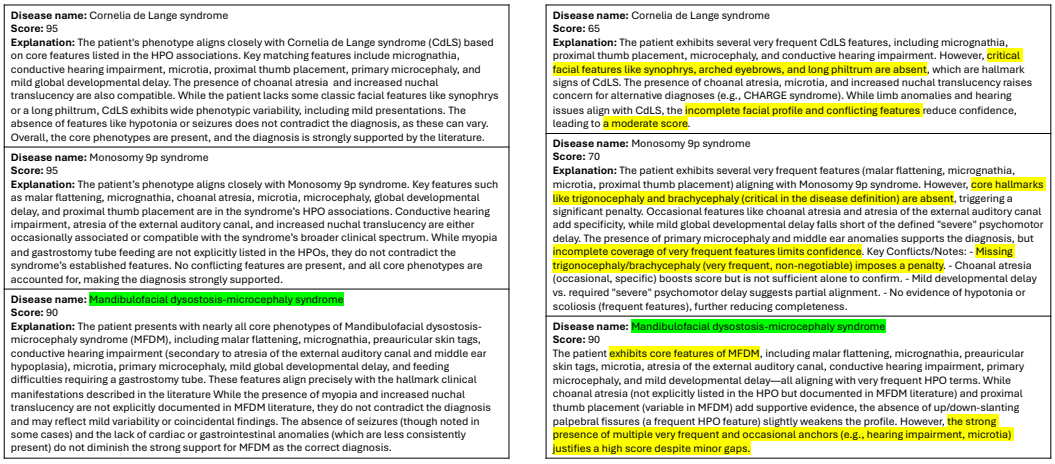
Table 2: **Ablation study** on the knowledge-driven self-evaluation and multi-evidence fusion of our method across three datasets.

	MME			LIRICAL			RAMEDIS		
Method	R@1	R@3	R@5	R@1	R@3	R@5	R@1	R@3	R@5
RADAR (Ours)	**57.50**	**72.50**	**82.50**	**42.97**	**63.24**	**74.05**	**41.66**	**68.26**	**78.52**
`RADAR-V1`	25.00	42.50	62.50	26.75	56.21	63.78	28.84	55.60	65.70
Δ (*vs.* Ours)	32.50 ↓	30.00 ↓	20.00 ↓	16.22 ↓	7.03 ↓	10.27 ↓	12.82 ↓	12.66 ↓	12.82 ↓
`RADAR-V2`	47.50	67.50	70.00	35.13	58.10	67.83	33.97	61.21	70.51
Δ (*vs.* Ours)	10.00 ↓	5.00 ↓	12.50 ↓	7.84 ↓	5.14 ↓	6.22 ↓	7.69 ↓	7.05 ↓	8.01 ↓
`RADAR-V3`	52.50	70.00	77.50	38.10	61.62	72.97	38.14	67.46	78.20
Δ (*vs.* Ours)	5.00 ↓	2.50 ↓	5.00 ↓	4.87 ↓	1.62 ↓	1.08 ↓	3.52 ↓	0.80 ↓	0.32 ↓

observe that each tool exhibits dataset-specific advantages: PhenoBrain performs best on MME, PubCaseFinder on LIRICAL, and the Doctor Agent on RAMEDIS. This performance variance demonstrates that relying on any single source of evidence is insufficient to address the inherent diversity of rare diseases. By synergistically integrating these disparate sources, RADAR significantly outperforms all individual pathways across all three benchmarks. Notably, RADAR improves R@1 from 22.50% to **57.50**% on MME, from 25.94% to **42.97**% on LIRICAL, and from 34.45% to **41.66**% on RAMEDIS. These gains demonstrate that RADAR effectively leverages the complementary strengths of each tool, mitigating the biases and coverage gaps inherent in single-evidence diagnostics.

Effectiveness of Knowledge-Driven Self-Evaluation. To investigate the effectiveness of the proposed knowledge-driven evaluation module, we construct two variants of RADAR, namely `RADAR-V1` and `RADAR-V2`. Specifically, `RADAR-V1` replaces the knowledge-driven evaluation with simple heuristic scoring rules, and `RADAR-V2` removes the chain-of-thought prompting during evaluation. As shown in Table 2, the performance gap between RADAR and `RADAR-V1` is the most pronounced. Specifically, the R@1 score drops sharply from **57.50**% to **25.00**% on MME. This significant degradation suggests that simple heuristic rules are insufficient for evaluating heterogeneous diagnostic outputs, particularly in challenging cases where the ground-truth disease shares overlapping clinical phenotypes with competing candidates (see Figure 5 (a)). It highlights that our knowledge-driven evaluation mechanism is essential for guiding the agent toward accurate diagnosis by effectively distinguishing subtle differences and filtering out misleading alternatives (see Figure 5 (b)). By comparing RADAR with `RADAR-V2`, we observe a consistent performance decline across three datasets. This result validates that explicit reasoning paths help the evaluator agent to better align the provided knowledge with the candidate answers, leading to more robust and calibrated assessment outcomes.

Effectiveness of Multi-Evidence Fusion. We further construct `RADAR-V3` to evaluate the effectiveness of the proposed multi-evidence fusion strategy. In this variant, the Aggregator Agent directly combines structured diagnostic outputs from different sources without performing the hierarchical fusion process. As shown in Table 2, `RADAR-V3` consistently underperforms our method across all datasets. This indicates that simple aggregation of multi-source evidence is insufficient for reliable



(a) RADAR-V1 (b) RADAR (Ours)

Figure 5: **Examples** of evaluation results from `RADAR-V1` and our method, where the ground-truth disease is highlighted.

diagnosis, as the ranking process becomes unstable when the evaluator assigns similar scores to different candidates or when different sources produce partially overlapping predictions. Effective multi-evidence reasoning requires not only collecting outputs from different sources, but also explicitly modeling their reliability, cross-source agreement, and justification during aggregation. These results validate the effectiveness of our proposed fusion strategy in producing more stable and reliable diagnostic decisions.

4.4 Effciency

To evaluate the practical deployment potential of our framework, we compare its computational efficiency against DeepRare. As shown in Table 3, RADAR achieves significantly lower inference time and context length compared to DeepRare. This improvement can be attributed to two main factors. First, DeepRare relies on an iterative reflection mechanism, which introduces additional reasoning steps and leads to substantially higher inference time. In contrast, RADAR removes this explicit reflection process, resulting in a more streamlined inference pipeline. Second, DeepRare adopts a tightly coupled diagnostic pipeline, where multiple sources of evidence are aggregated within a single context. This design leads to long context sequences during inference. By decoupling the diagnostic process into specialized workflows, RADAR reduces unnecessary context accumulation and significantly lowers token consumption. Overall, these designs make RADAR more efficient for real-world deployment while maintaining strong diagnostic performance.

Table 3: **Efficiency comparison** with DeepRare using inference time and context length.

Method	Time (s)	Context (token)
DeepRare	1986	3130
RADAR (Ours)	**1021**	**1897**

5 Conclusion

In this work, we presented RADAR, a rare disease diagnosis agent designed for small-scale underlying LLMs. The key idea is to decouple the diagnostic process into specialized workflows, enabling more structured handling of heterogeneous evidence from multiple sources. We further introduce a knowledge-driven self-evaluation mechanism to assess the reliability of different diagnostic outputs with interpretable rationales. Based on these evaluations, a multi-evidence fusion strategy is used to integrate multi-source evidence and produce a stable ranking of candidate diseases. Extensive experiments on three rare disease benchmarks show that RADAR consistently outperforms both commercial LLMs and recent agent-based systems. More importantly, it achieves substantial improvements over DeepRare with the same small-scale underlying LLM in both efficiency and recall, providing a practical framework for resource-constrained settings.

Limitation. RADAR focuses on phenotype-based diagnosis and is evaluated on HPO-based datasets. While this design ensures consistency with structured clinical representations, it relies on preprocessed phenotype inputs. For free-text phenotypes, similar to DeepRare, these inputs can be mapped to standardized HPO terms using LLM-based phenotype normalization, which can be integrated into our framework as a preprocessing step. In addition, incorporating genetic information is a promising direction for future work.

Acknowledgements

This work is supported by Guangdong Science and Technology Department (2024ZDZX2004), Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515140021), and the National Natural Science Foundation of China (Project No.82572383).

Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)

[2] Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., Hamosh, A.: Omim. org: Online mendelian inheritance in man (omim®), an online catalog of human genes and genetic disorders. Nucleic acids research **43**(D1), D789–D798 (2015)

[3] Anthropic: Claude sonnet. https://www.anthropic.com/claude/sonnet (2025), product webpage

[4] Chen, X., Mao, X., Guo, Q., Wang, L., Zhang, S., Chen, T.: 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 (2024)

[5] Collaco, B.G., Haider, S.A., Prabha, S., Gomez-Cabello, C.A., Genovese, A., Wood, N.G., Bagaria, S.P., Gopala, N., Tao, C., Forte, A.J.: The role of agentic artificial intelligence in healthcare: a scoping review. npj Digital Medicine (2026)

[6] Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)

[7] Deisseroth, C.A., Birgmeier, J., Bodle, E.E., Kohler, J.N., Matalon, D.R., Nazarenko, Y., Genetti, C.A., Brownstein, C.A., Schmitz-Abe, K., Schoch, K., et al.: Clinphen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis. Genetics in Medicine **21**(7), 1585–1593 (2019)

[8] Dou, C., Liu, C., Yang, F., Li, F., Jia, J., Chen, M., Ju, Q., Wang, S., Dang, S., Li, T., et al.: Baichuan-m2: Scaling medical capability with large verifier system. arXiv preprint arXiv:2509.02208 (2025)

[9] Fan, Z., Wei, L., Tang, J., Chen, W., Siyuan, W., Wei, Z., Huang, F.: Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. In: Proceedings of the 31st International Conference on Computational Linguistics. pp. 10183–10213 (2025)

[10] Fujiwara, T., Yamamoto, Y., Kim, J.D., Buske, O., Takagi, T.: Pubcasefinder: A case-report-based, phenotype-driven differential-diagnosis system for rare diseases. The American Journal of Human Genetics **103**(3), 389–399 (2018)

[11] GLM-5-Team, et al.: Glm-5: from vibe coding to agentic engineering (2026)

[12] Gou, Z., Shao, Z., Gong, Y., Shen, Y., Yang, Y., Huang, M., Duan, N., Chen, W.: Tora: A tool-integrated reasoning agent for mathematical problem solving. arXiv preprint arXiv:2309.17452 (2023)

[13] Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)

[14] Jin, Q., Kim, W., Chen, Q., Comeau, D.C., Yeganova, L., Wilbur, W.J., Lu, Z.: Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. Bioinformatics **39**(11), btad651 (2023)

[15] Jonker, A.H., Cavaller-Bellaubi, M., Nishimura, Y., Pearce, D.A.: Access in the rare diseases landscape. The Lancet Global Health **12**(10), e1587 (2024)

[16] Kim, Y., Park, C., Jeong, H., Chan, Y.S., Xu, X., McDuff, D., Lee, H., Ghassemi, M., Breazeal, C., Park, H.W.: Mdagents: An adaptive collaboration of llms for medical decision-making. Advances in Neural Information Processing Systems **37**, 79410–79452 (2024)

[17] Köhler, S., Schulz, M.H., Krawitz, P., Bauer, S., Dölken, S., Ott, C.E., Mundlos, C., Horn, D., Mundlos, S., Robinson, P.N.: Clinical diagnostics in human genetics with semantic similarity searches in ontologies. The American Journal of Human Genetics **85**(4), 457–464 (2009)

[18] Lee, J., Liu, C., Kim, J., Chen, Z., Sun, Y., Rogers, J.R., Chung, W.K., Weng, C.: Deep learning for rare disease: A scoping review. Journal of biomedical informatics **135**, 104227 (2022)

[19] Li, B., Yan, T., Pan, Y., Luo, J., Ji, R., Ding, J., Xu, Z., Liu, S., Dong, H., Lin, Z., et al.: Mmedagent: Learning to use medical tools with multi-modal agent. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 8745–8760 (2024)

[20] Li, X., Dong, G., Jin, J., Zhang, Y., Zhou, Y., Zhu, Y., Zhang, P., Dou, Z.: Search-o1: Agentic search-enhanced large reasoning models. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 5420–5438 (2025)

[21] Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al.: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024)

[22] Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al.: Self-refine: Iterative refinement with self-feedback. Advances in neural information processing systems **36**, 46534–46594 (2023)

[23] Mao, X., Huang, Y., Jin, Y., Wang, L., Chen, X., Liu, H., Yang, X., Xu, H., Luan, X., Xiao, Y., et al.: A phenotype-based ai pipeline outperforms human experts in differentially diagnosing rare diseases using ehrs. NPJ Digital Medicine **8**(1), 68 (2025)

[24] Moonshot AI: Kimi k2.6: Advancing open-source coding. https://www.kimi.com/ai-models/kimi-k2-6 (2026), technical blog

[25] Robinson, P.N., Ravanmehr, V., Jacobsen, J.O., Danis, D., Zhang, X.A., Carmody, L.C., Gargano, M.A., Thaxton, C.L., Karlebach, G., Reese, J., et al.: Interpretable clinical genomics with a likelihood ratio paradigm. The American Journal of Human Genetics **107**(3), 403–417 (2020)

[26] Schieppati, A., Henter, J.I., Daina, E., Aperia, A.: Why rare diseases are an important medical and social issue. The Lancet **371**(9629), 2039–2041 (2008)

[27] Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)

[28] Shang, J., Xiao, C., Ma, T., Li, H., Sun, J.: Gamenet: Graph augmented memory networks for recommending medication combination. In: proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 1126–1133 (2019)

[29] Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., Yao, S.: Reflexion: Language agents with verbal reinforcement learning. Advances in neural information processing systems **36**, 8634–8652 (2023)

[30] Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)

[31] Smedley, D., Jacobsen, J.O., Jäger, M., Köhler, S., Holtgrewe, M., Schubach, M., Siragusa, E., Zemojtel, T., Buske, O.J., Washington, N.L., et al.: Next-generation diagnostics and disease-gene discovery with the exomiser. Nature protocols **10**(12), 2004–2015 (2015)

[32] Tang, X., Zou, A., Zhang, Z., Li, Z., Zhao, Y., Zhang, X., Cohan, A., Gerstein, M.: Medagents: Large language models as collaborators for zero-shot medical reasoning. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 599–621 (2024)

[33] Valdez, R., Ouyang, L., Bolen, J.: Public health and rare diseases: oxymoron no more. Preventing chronic disease **13**, E05 (2016)

[34] Wang, W., Ma, Z., Wang, Z., Wu, C., Ji, J., Chen, W., Li, X., Yuan, Y.: A survey of llm-based agents in medicine: How far are we from baymax? Findings of the Association for Computational Linguistics: ACL 2025 pp. 10345–10359 (2025)

[35] Wang, Z., Wu, J., Cai, L., Low, C.H., Yang, X., Li, Q., Jin, Y.: Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. arXiv preprint arXiv:2503.18968 (2025)

[36] Weinreich, S.S., Mangon, R., Sikkens, J., Teeuw, M.E., Cornel, M.: Orphanet: a european database for rare diseases. Nederlands tijdschrift voor geneeskunde **152**(9), 518–519 (2008)

[37] Wölflein, G., Ferber, D., Truhn, D., Arandjelovic, O., Kather, J.N.: Llm agents making agent tools. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 26092–26130 (2025)

[38] Wu, Z., Wang, Z., Xu, X., Lu, J., Yan, H.: Embodied task planning with large language models. arXiv preprint arXiv:2307.01848 (2023)

[39] Xie, Y., Kawaguchi, K., Zhao, Y., Zhao, J.X., Kan, M.Y., He, J., Xie, M.: Self-evaluation guided beam search for reasoning. Advances in Neural Information Processing Systems **36**, 41618–41650 (2023)

[40] Xu, W., Liang, Z., Mei, K., Gao, H., Tan, J., Zhang, Y.: A-mem: Agentic memory for llm agents. arXiv preprint arXiv:2502.12110 (2025)

[41] Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)

[42] Yang, C., Xiao, C., Glass, L., Sun, J.: Change matters: Medication change prediction with recurrent residual networks. arXiv preprint arXiv:2105.01876 (2021)

[43] Yang, C., Jin, H., Yu, X., Wang, Z., Liu, Y., Fan, F., Lei, D., Jia, G., Wang, C., Ge, R.: Lungnoduleagent: A collaborative multi-agent system for precision diagnosis of lung nodules. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 29793–29801 (2026)

[44] Yang, H., Robinson, P.N., Wang, K.: Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. Nature methods **12**(9), 841–843 (2015)

[45] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K.R., Cao, Y.: React: Synergizing reasoning and acting in language models. In: The eleventh international conference on learning representations (2023)

[46] Zhai, W., Huang, X., Shen, N., Zhu, S.: Phen2disease: a phenotype-driven model for disease and gene prioritization by bidirectional maximum matching semantic similarities. Briefings in Bioinformatics **24**(4), bbad172 (2023)

[47] Zhao, L., Liu, S., Xin, T., Tan, J., Wang, X., Li, Y., Bian, Z., Chen, Y., Kong, F., Bian, J., et al.: Ai agent in healthcare: applications, evaluations, and future directions. npj Artificial Intelligence **2**(1), 31 (2026)

[48] Zhao, W., Wu, C., Fan, Y., Qiu, P., Zhang, X., Sun, Y., Zhou, X., Zhang, S., Peng, Y., Wang, Y., et al.: An agentic system for rare disease diagnosis with traceable reasoning. Nature pp. 1–10 (2026)

[49] Zhong, W., Guo, L., Gao, Q., Ye, H., Wang, Y.: Memorybank: Enhancing large language models with long-term memory. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 19724–19731 (2024)

[50] Zou, K., Goh, J.H.L., Yang, G.D., Srinivasan, S., Chen, Q., Keane, P.A., Tham, Y.C.: Agentic artificial intelligence in eye care: is clinical autonomy finally within reach? The Lancet Digital Health **8**(3) (2026)

A Prompts

We present the main prompts in our method and experiments as follows:

Prompt 1: Diagnosis Prompt for Doctor Agent

You have access to the following context:
- **Online knowledge** (with titles and URLs): {`web knowledge`}
- **Similar cases** (detailed): {`similar cases`}
- **Patient's phenotype**: {`patient phenotype`}
- - -
Based on the above and your knowledge, enumerate the **top 5 most likely rare disease diagnoses** for this patient.
For each diagnosis, use the following format:
<|begin|>DIAGNOSIS NAME<|end|> (Rank #X/5)

Prompt 2: Judgment Prompt for Evaluator Agent

You are a medical expert specialized in rare disease diagnosis. Your task is to evaluate a candidate diagnosis.
Frequency-Based Logic (The Algorithm)
You must treat the HPO frequency tiers with different logical weights:
1. **Very Frequent (99-80%) - "The Threshold"**
- **Role**: These are the Pillars.
- **Logic**: Absence of these features leads to a **proportional penalty**, not a hard exclusion.
- Penalty should scale with missing coverage (e.g., more missing → larger penalty).
- **Important**: Absence reduces confidence but does NOT rule out the disease.
- **Inference**: High match in this category indicates the diagnosis is "Possible," but not sufficient alone.
2. **Frequent (79-30%) - "The Profile"**
- **Role**: These build the clinical profile.
- **Logic**: Matches provide steady accumulation of confidence and help distinguish similar diseases.
3. **Occasional (29-5%) - "The Clincher"**
- **Role**: These are specific supporting clues.
- **Logic**: Presence provides a **moderate bonus** if consistent with the disease.
- Absence should NOT be penalized.
- - -
Additional Rule: Conflict Awareness (Critical)
- **Definition of a Conflict:** A phenotype is considered a conflict only if it is:
1. Logically or biologically opposite (e.g., high vs low, increased vs decreased)
2. Mutually exclusive with defining features of the disease
3. Represents a state incompatible with the disease
- Apply penalties:
- Major conflict (true contradiction): **-30 points each**
- Minor inconsistency (atypical but possible): **0 to -3 points**, optional
- **Important:**
- Conflicts are more decisive than missing features
- Do NOT treat unrelated or rare-but-possible features as conflicts
- - -
Evaluation Steps
Step 1: The Definition Check (Priority 1)
Analyze the **Disease Definition**. Identify the 2–3 key hallmarks.
- Evaluate whether they are: present / partially present / absent
- If absent: apply a **moderate penalty (-10 to -15)** (NOT a hard cap)
- - -

Step 2: Weighted Coverage Analysis
- Estimate coverage for: Very Frequent features / Frequent features
- Penalize missing Very Frequent features **proportionally**, not absolutely
- Down-weight generic HPOs (e.g., global developmental delay) if they dominate matches
- - -
Step 3: Clinical Signature Synthesis
- Evaluate whether the combination of features forms a recognisable disease pattern: strong match / partial / weak
- - -
Step 4: Conflict Check (Mandatory)
- Identify any true conflicts (per definition above)
- Apply penalties as described in "Conflict Awareness"
- - -
Scoring Guidance (Important)
- Start from an overall clinical impression based on:
  - coverage (Very Frequent + Frequent)
  - signature matching
- Then adjust:
  - subtract moderate penalty for missing key features
  - subtract stronger penalty for conflicts
- Avoid hard thresholds or caps unless evidence is extremely inconsistent
- - -
Input Data
- **Candidate Diagnosis**: {`candidate disease`}
- **Disease Knowledge**: {`disease knowledge`}
- **Patient Phenotype**: {`patient phenotype`}
- - -
Return your answer strictly in the following format:
SCORE: [0-100]
EXPLANATION: [Provide a concise explanation in 3-5 sentences. Focus on whether core phenotypes are present, partially present, or absent, and note any conflicting features.]

Prompt 3: Fusion Prompt for Aggregator Agent

You are a medical decision fusion system
Below are structured outputs from multiple independent diagnostic tools:
Each tool provides structured entries in the following format:
{
"disease_name": string,
"judge_result": integer (0-100),
"judgments": detailed clinical explanation
}
Important:
- same disease may appear in multiple tools.
- judgments may provide additional reasoning to help rank diseases with equal scores.
==========
TOOL OUTPUTS
==========
PHENOBRAIN RESULTS: {`phenobrain result`}
PUBCASEFINDER RESULTS: {`pubcasefinder result`}
DOCTOR AGENT RESULTS: {`doctor agent result`}
==========
FUSION PROCESS (Follow strictly)
==========
STEP 1:
Identify all unique diseases.
STEP 2:
For each disease:
- Record its judge_result.
- Count how many tools predicted it (support count).
STEP 3: Ranking priority:
1. Higher judge_result ranks higher.
2. If judge_result is equal, the disease predicted by MORE tools must rank higher.
3. If judge_result AND support counts are equal, use ‘judgments‘ to evaluate which disease is better supported by the patient’s phenotype
4. Do not modify scores.
5. Do not introduce new diseases.
==========
IMPORTANT RULES:
- Output exactly 5 unique diseases.
- No duplicates.
- Preserve disease names exactly.
- Only use diseases present in the tool outputs.
==========
FINAL Output:
<|begin|>DIAGNOSIS NAME<|end|> (Rank #X/5)
SCORE: <judge_result>
EXPLANATION: <judgments>

Prompt 4: Judge Prompt for GPT-5.1

You are a specialist in the field of rare diseases. I will now give you five predicted diseases if the predicted diagnosis is in the standard diagnosis. Please output the predicted rank, otherwise output "No", only output "No" or "1-5" numbers, if the predicted disease has multiple conditions, only output the top rank. Output only "No" or one number, no additional output.
Predicted diseases: {`predicted diseases`}
Standard diagnosis: {`ground-truth disease`}

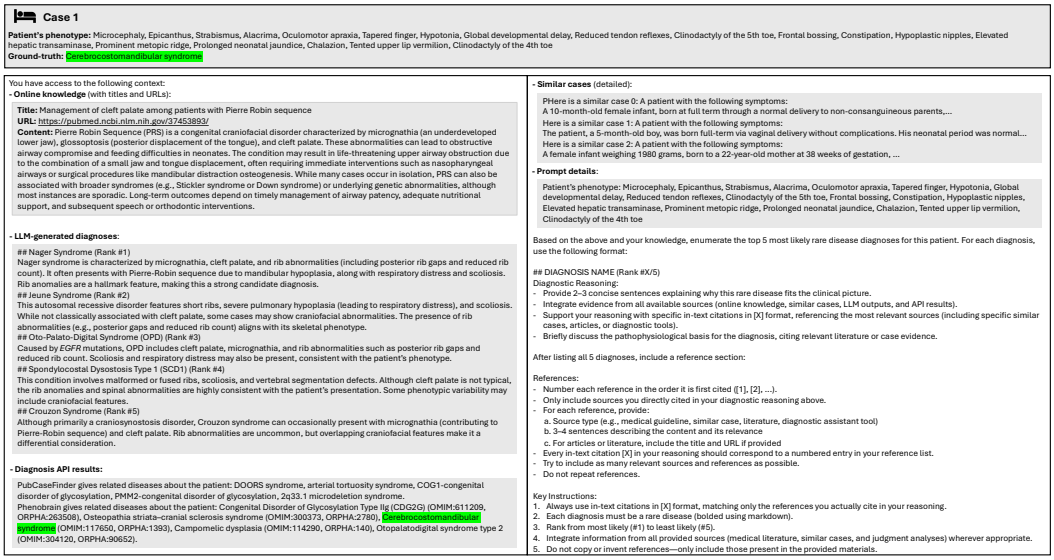


Figure 6: **Illustration of** (a) a patient case (top) and (b) contextual content in DeepRare (bottom).

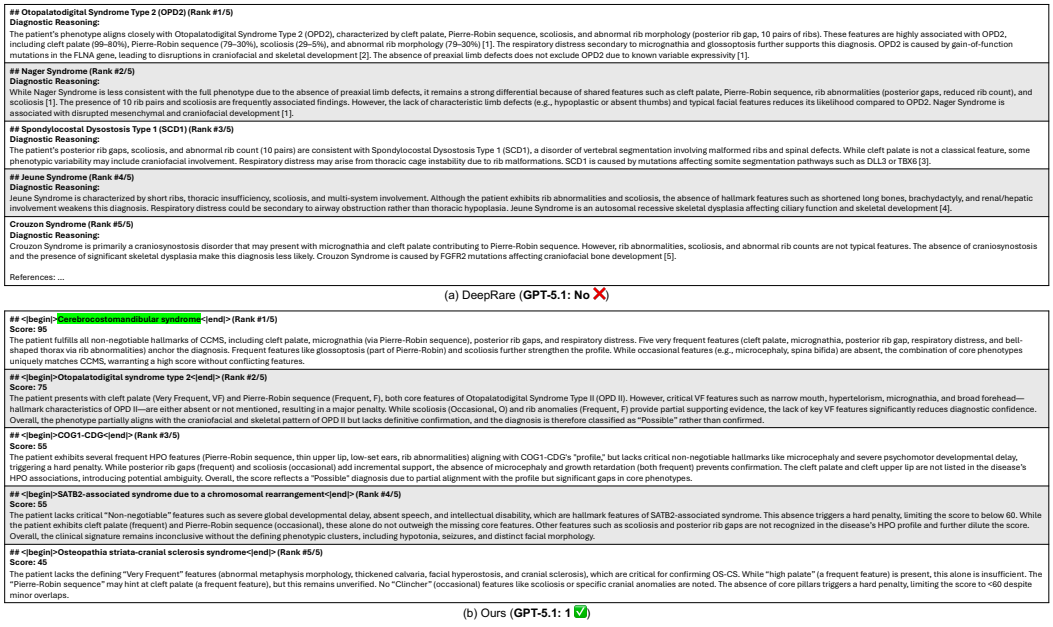


Figure 7: **Case study comparison** between (a) DeepRare and (b) ours.

B Case Study

In this section, we conduct a case study to gain a deeper understanding of DeepRare and our method. Figure 6 illustrates a patient case and the contextual content in DeepRare, and Figure 7 shows the diagnosis results of DeepRare and our method. DeepRare directly incorporates all evidence into the context, including outputs from multiple diagnostic tools, similar cases, and external API-retrieved knowledge. This design implicitly requires the underlying LLM to simultaneously perform multi-source evidence aggregation, conflict resolution, and complex reasoning over heterogeneous information. Such a monolithic pipeline entangles multiple reasoning tasks within a single model, imposing a substantial cognitive burden. Small-scale models struggle to effectively handle long contexts, leading to insufficient information filtering and prioritization. As shown in Figures 6 and 7 (a), the output from the diagnostic tool phenobarin contains correct information consistent with

the ground truth, yet it is ignored by the model. Moreover, the diagnostic reasoning indicates that its decision is primarily driven by simple phenotype matching, without effectively capturing the varying importance of different phenotypic features.

In contrast, our method decouples heterogeneous evidence into separate workflows, enabling more structured and specialized processing. The introduced knowledge-driven self-evaluation mechanism enables the assessment of the reliability of each diagnostic signal. Based on the structured evaluation results, we can effectively select reliable diagnostic evidence and resolve conflicts among different diagnostic tools, as shown in Figure 7 (b).

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction accurately reflect the contributions and scope of the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: Section 5.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper focuses on empirical methods and does not present formal theoretical results.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4.1.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The datasets used in this work are publicly available. The code cannot be publicly released due to company confidentiality policies and restrictions. We provide sufficient details of the experimental settings, implementation, and evaluation protocols to facilitate understanding and reproducibility of the main experimental results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The experiments are conducted under fixed settings and evaluated using standard benchmarks.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: All experimental details are presented in Section 4.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research presented in this paper complies with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper does not explicitly discuss broader societal impacts. However, the proposed method is intended for rare disease diagnosis, which may have positive societal benefits.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper poses no such risks.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets, models, and codebases used in this work are properly credited through citations to their original sources. The datasets are publicly available and used in accordance with their respective licenses and terms of use.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not introduce new datasets or external assets. The proposed method is fully described in the paper, including all components of the framework. The implementation will be released upon acceptance to facilitate reproducibility.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: We only used LLMs for minor tasks such as grammar checking and expression polishing. The LLMs did not contribute to research ideation, experimental design, analysis, or drafting of technical content. All scientific claims, experimental results, and writing remain the sole responsibility of the authors.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.