

Ontology-guided multi-agent reasoning for explainable differential diagnosis

Roxana Ramona Szomiu, George Ursu, Adrian Groza

Artificial Intelligence Research Institute AIRi@UTCN

Department of Computer Science

Technical University of Cluj-Napoca

Cluj-Napoca, Romania

Email: Roxana.Szomiu@cs.utcluj.ro, Ursu.Da.George@student.utcluj.ro, Adrian.Groza@cs.utcluj.ro

Abstract—Effective differential diagnosis depends on well-structured and unambiguous patient case descriptions, since incomplete or ambiguous clinical information readily leads to diagnostic errors and inconsistent reasoning. Large Language Models (LLMs) have recently shown promising capabilities in medical question answering and clinical reasoning; however, even when achieving competitive performance on medical benchmarks, they often rely on pattern recognition learned from training data rather than grounded clinical evidence, and can generate plausible yet unsupported diagnostic conclusions. Their reasoning remains largely opaque, making it difficult for clinicians to assess the evidence and rationale behind a given recommendation.

To address this gap, we propose an ontology-guided, multi-agent differential diagnosis framework that replicates the iterative reasoning process of physicians. The framework grounds diagnostic inference in a structured, expert-curated medical knowledge base, enabling interpretation of symptoms, clinical findings, or disease relationships, while decomposing diagnosis into explicit, auditable reasoning steps. This ontology-based grounding constrains inference to evidence-backed associations, reducing hallucination and improving diagnostic transparency, traceability, and consistency. The system then identifies the most discriminative missing symptoms and generates targeted follow-up questions to reduce uncertainty through a process that mirrors to clinical inquiry. Together, these components improve the reliability, explainability, and diagnostic accuracy of AI-assisted differential diagnosis.

Index Terms—differential diagnosis, multi-agent systems, LLMs, ontology-based reasoning, explainable AI

I. INTRODUCTION

Differential diagnosis represents an important task in the medical domain that is reached in an iterative, step-by-step process used to distinguish between two or more diseases that have similar symptoms. Initial patient notes are often incomplete and may omit key details or contain uncertain information. This makes it difficult to achieve reliable diagnostic decision-making without further investigation.

Early evaluations of LLMs in healthcare formulated diagnostic reasoning as a *single-turn question-answering task*, assuming that all the information required for diagnosis and treatment selection was available at the time of prediction [1].

The transition from static question-answering systems [2] to interactive multi-turn diagnostic dialogues is crucial for clinical decision-making. To reduce uncertainty and to progressively improve diagnostic accuracy, physicians need to

iteratively collect relevant information through targeted follow-up questions. This highlights that clinical diagnosis is better understood as an interactive, multi-turn process rather than a single-step inference task.

Motivated by the demonstrated potential of LLMs to improve differential diagnosis and support clinical decision-making [3], we propose an ontology-based multi-agent system that replicates the iterative clinical reasoning and that can assist clinicians in the process of establishing a correct diagnosis.

The multi-agent neuro-symbolic framework for explainable differential diagnosis integrates LLM-based language understanding with ontology-driven symbolic reasoning. The framework is organized in 2 modules and adopts a task-oriented multi-agent architecture, where specialized LLM agents collaborate to interpret clinical narratives, extract structured symptoms, and manage patient interactions.

The symbolic reasoning module, grounded in medical ontologies (SYMP [4], DOID [5], [6], and HPO [7]), performs transparent disease scoring and differential diagnosis. The module queries the underlying medical knowledge graph to identify the most informative symptoms that distinguish among competing diagnoses. Based on this analysis, the dialogue-generator agent autonomously formulates targeted Yes/No follow-up questions that efficiently reduce diagnostic uncertainty, and maximize diagnostic information gain. Rather than relying on a single prediction from a fixed set of symptoms, the framework incrementally refines the differential diagnosis as new clinical evidence becomes available.

By combining LLMs with ontology reasoning, the proposed multi-agent system supports an adaptive and explainable diagnostic workflow. Explainability is achieved through the traceability of the decision-making process: at each phase, mathematically defined scoring functions rank candidate diagnoses based on extracted clinical evidence and structured ontology relationships. Instead of generating only the final diagnosis, the system provides a transparent, traceable reasoning process in which each refinement of the differential diagnosis can be traced back to the contributing symptoms, clinical findings, and mathematical scores. This allows clinicians to interpret both the final diagnosis and the reasoning behind it, making the diagnostic process interpretable, transparent, and clinically meaningful.

II. RELATED WORK

Singhal et al. [8] introduced Med-PaLM2, a medical LLM that achieved expert-level performance on several clinical question-answering benchmarks, including 86.5% accuracy on MedQA, representing a more than 19% improvement over Med-PaLM [9]. However, the authors identified an important limitation of the single-turn question-answering paradigm: even fully specified symptom descriptions rarely determine a diagnosis completely. Since patients present partial symptoms descriptions, the doctors have to actively gather additional information through targeted follow-up questions.

Muntean et al. [10] demonstrate that Med-PaLM2 can provide accurate and clinically appropriate responses for patient education tasks, to ophthalmology-related questions. However, the interaction is limited to a single-turn exchange, which prevents the model from clarifying ambiguities or refining differential diagnoses through iterative dialogue. This limitation has motivated a shift from static question-answering benchmarks to interactive diagnostic settings, where diagnoses are progressively constructed through iterative question-answer interactions between the doctor and the patient.

The authors [11] introduced the Q4Dx benchmark to evaluate the strategic inquiry capability of a diagnostic system rather than its static medical knowledge. Their starting assumption is that real doctor-patient encounters begin with incomplete, uncertain, and noisy descriptions expressed in non-specialist language. Doctors must then recover the missing critical information through an adaptive sequence of focused questions, each question based on the patient's previous responses.

The synthetic patient notes were generated from curated symptom-diseases pairings (SymbiPredict dataset [12]), with different levels of symptom completeness (100%, 80% and 50%) together with the ground truth diagnosis. To assess how effectively systems reduce diagnostic uncertainty over multiple turns, the patient-physician interactions are simulated with GPT-4.1 and GPT-4o-mini agents. The study reveals that while partial information significantly reduces single-shot diagnostic performance, interactive questioning can recover much of this loss, underscoring the importance of dialogue-based reasoning in clinical diagnosis. The results obtained by the baseline models show an accuracy degradation from 51% to 44%, and then to 37%, as the level of information disclosure decreases from 100% to 80% then to 50%.

Recent work on interactive clinical reasoning introduces complementary benchmarks and frameworks that together advance multi-turn diagnostic modeling. MedKGI [13] improves diagnostic dialogue by grounding reasoning in medical knowledge graphs, selecting questions based on information gain, and maintaining structured state tracking, leading to more efficient and coherent interactions. MedPI [14] evaluates nine flagship models across 366 AI patients and 7,100 conversations, and shows that despite strong performance on static benchmarks, LLMs still struggle with full patient-clinician conversations, particularly in differential diagnosis. While MINT [15] reveals issues such as premature diagnostic

commitment, sensitivity to misleading evidence, and limited self-correction. They show that delayed questioning and structured evidence presentation significantly improve accuracy by up to 62.6%. Overall, these studies demonstrate that effective clinical diagnosis requires grounded, multi-turn, and temporally aware reasoning rather than single-step prediction.

Gond et al. [16] introduce MedDialogRubrics, a benchmark of 5,200 synthetic patient cases and over 60,000 expert-refined rubrics for multi-turn clinical consultations, and show that even strong models (e.g., Gemini-2.5-Pro) miss nearly half of the must-ask questions. Their analysis also reveals that diagnostic performance peaks within a limited interaction window (around 10-11 turns) and degrades when conversations become longer. This findings directly motivate a design choice in the our system, whose follow-up interviewer agent has maximum limit of 11 questions: accuracy is maximized within this interaction length, while additional questioning beyond this point reduces diagnostic alignment and unnecessarily prolongs the consultation.

MedAgents [17] is one of the first multi-agent collaboration frameworks designed specifically for the medical domain. Evaluated on nine multiple-choice datasets, MedAgents reaches an average accuracy of 86.7% with GPT-4, slightly above a stronger few-shot CoT (Chain-of-Thought) baseline despite using no in-context examples.

III. METHODOLOGY

MedMAS-DDX is a multi-agent system that treats differential diagnosis as a multi-step process, not a single inference. It translates patient's free-text complaint to a ranked, explained list of probable diseases through a sequence of transformations: unstructured language becomes structured symptoms, symptoms become candidate diseases, and candidate diseases are organized into an initial ranking representing the differential diagnosis. That ranking is then sharpened by a short exchange of questions, and the answers are fed back into the scoring to produce the final diagnosis.

Two principles govern every step. First, responsibility is split by reliability: language models handle only natural language understanding and generation, while all ranking-related computations are done by deterministic code, ensuring that diagnosis does not depend on model-based calculations. Second, all medical knowledge is stored in an external curated source and is explicitly queried during the process, rather than being learned from the model's parameters.

MedMAS-DDX realizes the preceding principles as a neuro-symbolic pipeline in which a small set of LLM-based agents is embedded within, and coordinated by, a deterministic control layer (Figure 1). The multi-agent system is implemented using the CrewAI framework, which manages agent definition and execution. Each agent is designed as a distinct role-goal-backstory persona operating on the same underlying LLM (accessed via the LiteLLM), with specialized rather than interchangeable functions.

The system consists of four agents: (i) a symptom extraction agent (`symptom_extractor`), (ii) an interview-

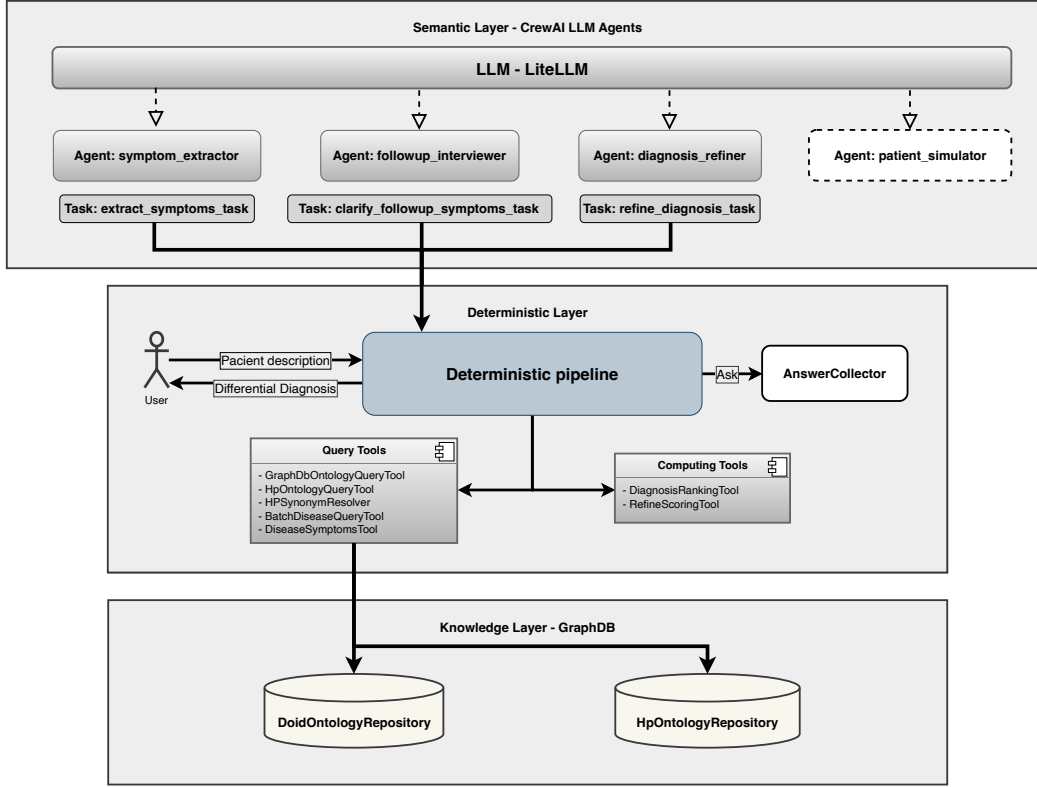


Fig. 1: System components and their interactions.

ing agent (followup_interviewer), (iii) an explanatory agent (diagnosis_refiner), (iv) a patient agent (patient_simulator) used only to drive automated evaluation and never contributes to a live diagnosis. The first three agents participate in the diagnostic-making process, while the last agent simulates the patient:

The multi-agent system is characterized by high control over execution, predictable processing paths, making it well-suited for strict business logic. In this architecture, agents do not communicate directly; instead, coordination is fully mediated by deterministic Python code. This control layer validates and reshapes the intermediate representations, ensuring that all interactions are explicitly structured and mediated by code rather than unconstrained agent-to-agent dialog.

The diagnostic-making pipeline allocates specific tasks to three active agents during a live diagnosis, while a fourth persona is reserved exclusively for methodological validation.

A. Symptom extractor agent

The symptom extractor agent, an LLM-based component without external tool access, constitutes the first step of the pipeline by parsing the patient's initial, natural-language clinical note - often incomplete and informal - and transforming it into a structured intermediate representation of symptoms. For each identified symptom, it generates a canonical medical term, alternative phrasings, and optional contextual details

such as duration or location, while keeping descriptive information separate from standardized terminology to ensure consistency and support ontology matching in the next step.

Our solution follows the methodology of extracting information from unstructured clinical notes and map them to corresponding ontology terms [18], [19]. According to this, in the second phase of the pipeline, the linking symptoms to diseases process occurs in two stages-*resolution* and *retrieval*-and constitutes the first interaction with external knowledge. In the *resolution stage*, each symptom is deterministically mapped to a standard medical concept through an ordered matching process that first attempts the canonical term, then alternative phrasings, and finally a secondary synonym vocabulary. Unresolved symptoms are retained for traceability, but excluded from further processing to avoid affecting the following inference steps.

In the *retrieval stage*, the resolved concepts are jointly used to query a curated knowledge graph of expert-defined symptom-disease relationships, enabling a single consolidated lookup rather than per-symptom queries. Diseases are included in the candidate set only if explicitly linked to at least one patient symptom, ensuring that all associations are grounded in verifiable expert knowledge. Each candidate disease is propagated together with the symptoms that triggered its retrieval.

All medical knowledge consulted during the pipeline is

stored in a GraphDB¹ triple store and is obtained exclusively through predefined SPARQL queries. This way, no disease-symptom relation is ever recalled from the LLM’s parameters. The knowledge base is partitioned into two repositories: one holds the Symptom Ontology (SYMP) and the Human Disease Ontology (DOID), while the second holds the Human Phenotype Ontology (HPO) and serves as a synonym bridge during vocabulary resolution. As queries are executed in Python rather than by LLM-based agents, knowledge access remains uniform, inspectable, and model-independent.

The retrieved candidate diseases, together with their associated symptoms, are forwarded to the third phase of the pipeline: the ranking phase. Candidate diseases are deterministically ranked based on symptom matches, where rarer and more informative symptoms receive higher weights. The resulting scores produce the initial differential diagnosis.

$$specificity_weight(s) = \frac{1}{\log_2(1 + N_d(s))} \quad (1)$$

Formula 1 encodes a key idea that rare symptoms are more informative than common symptoms. Let $N_d(s)$ denote the number of diseases associated with symptom s in the knowledge graph, and the logarithm reduces the impact of very large disease counts. If a symptom s appears in many diseases, then $N_d(s)$ is large, the denominator is large so weight becomes small (low specificity). If a symptom s appears in few diseases, then $N_d(s)$ is small, the denominator is small so weight becomes large (high specificity).

For each candidate disease d , a raw diagnostic score is computed by summing the diagnostic weights of all matched symptoms:

$$raw_score(d) = \sum_{s \in matched(d)} specificity_weight(s) \quad (2)$$

A *coverage factor* measures the proportion of the patient’s mapped symptoms explained by each candidate disease d :

$$coverage(d) = \frac{match_count(d)}{total_mapped_symptoms}. \quad (3)$$

Here the $match_count(d)$ is the number of patient symptoms matched by disease d , and $total_mapped_symptoms$ is the total number of successfully mapped symptoms. The *coverage factor* rewards diseases that account for a larger fraction of the reported symptoms, complementing the raw score based on symptom informativeness.

The raw score 2 and the coverage factor 3 are then combined into a final *composite score*:

$$comp_score(d) = raw_score(d) \times (0.5 + 0.5 \times coverage(d)) \quad (4)$$

The *composite score* combines two complementary aspects of diagnostic quality: the strength of supporting evidence and the completeness of symptom coverage. Here, $raw_score(d)$ measures the total diagnostic evidence for disease d , computed from the weighted contributions of matched symptoms. In

contrast, $coverage(d)$ measures how much of the patient’s symptom set is explained by the disease d .

The term $(0.5 + 0.5 \cdot coverage(d))$ defines a bounded scaling factor in $[0.5, 1]$ that adjusts rather than dominates the raw score. It ensures that coverage acts as a soft adjustment to the diagnostic evidence: when $(coverage(d)=0)$, the score is reduced but not eliminated (scale $(=0.5)$), while full coverage preserves the raw score (scale $(=1)$). This design prevents excessive penalization of low-coverage candidates while maintaining stable ranking behavior and balancing evidence strength with explanatory completeness.

An inclusive threshold then keeps every candidate that is plausibly relevant: those that reach a small fraction of the leading score, or that match at least two symptoms. The threshold is set to a low value on purpose, because a candidate discarded here cannot be recovered later. At this stage, a single remaining candidate is returned directly; otherwise, the process proceeds to questioning.

B. Follow-up interviewer agent

When multiple candidate diseases remain, the system enters a differentiation phase. The characteristic symptom profiles of the leading candidates are retrieved from the knowledge source and combined with the current ranking and reported symptoms to form an enriched state. The follow-up interviewer agent is responsible for resolving ambiguities and collecting missing clinical information.

Leveraging the enriched context, the LLM-based interviewer agent selects the most informative patient-observable symptoms by removing non-discriminative, or already reported, and generates targeted Yes/No follow-up questions to refine the clinical picture, without direct tool access and fully dependent on the state representation provided by the orchestration layer. To ensure a limited set of questions that support the differential diagnosis process, the agent keeps only symptoms that the patient can report and excludes those that require clinical examination or diagnostic testing.

At this stage, we have a human-in-the-loop interaction point in the system. After the system generates follow-up questions, the process stops temporarily to allow the patient (or user) to respond. The responses are constrained to a simple, structured format-“yes”, “no”, or “unsure”-which makes the input easy to process deterministically and avoids ambiguity in following reasoning steps. The mechanism that collects the answers is kept separate from the diagnostic logic, so the same pipeline runs unchanged whether answers come from a command line, a web interface, or an automated non-interactive source. This step turns a one-shot classification into a consultation. The system has formed the initial hypotheses and identified the most informative missing facts, and it obtains exactly those before concluding, much as a clinician narrows the possibilities through targeted questioning.

C. Patient simulator agent

To ensure rigorous and automated pipeline evaluation, a fourth persona is employed exclusively as a testing mecha-

¹<https://graphdb.ontotext.com/>

nism. The patient agent receives a fully specified case description and generates responses to the questions posed by the `followup_interviewer`, based only on the information contained in the current case. This agent operates entirely outside the production environment, with no external tool access and no involvement in live diagnostic sessions, serving exclusively for system validation purposes.

D. Diagnosis refiner agent

The diagnosis refiner agent serves as the final analytical component of the live diagnostic process, synthesizing all iteratively collected findings into a coherent medical conclusion. It takes as input the confirmed patient symptoms and the exact disease - symptom relations retrieved from the GraphDB triple store by the Python controller and produce a final diagnosis, along with a clear, evidence-based explanation of the clinical findings.

The patient agent’s answers are reintroduced into the pipeline and used to update the differential diagnosis through deterministic scoring. For each queried symptom s , a *discriminative weight* is computed as:

$$\text{discriminative_weight}(s) = 1 - \frac{\text{diseases_with_symptom}(s)}{n_{\text{profiled}}} \quad (5)$$

The $\text{diseases_with_symptom}(s)$ denotes the number of profiled candidate diseases that present symptom s , and n_{profiled} is the total number of profiled candidate diseases. This weight quantifies the ability of a symptom to distinguish among the remaining candidates: symptoms shared by many diseases receive lower weights, whereas symptoms present in only a few candidates receive higher weights. Consequently, responses to highly discriminative symptoms have a greater influence on the final ranking, while common symptoms contribute less because they provide limited additional diagnostic information.

For each disease, only the relevant questions are considered, that is, those whose symptom lies in its profile. The *discriminative weights* of the confirmed symptoms and of the denied symptoms are summed and normalised by the number of relevant questions into a *support term* and a *conflict term*. These adjust the prior *composite score* 4 through a bounded multiplicative update:

$$\text{final_score}(d) = \text{composite_score}(d) \times \frac{1}{(1 + \alpha \text{support_norm}(d) - \beta \text{conflict_norm}(d))} \quad (6)$$

where confirmations raise the score and denials lower it in proportion to discriminativeness. The parameters α and β control how much the update changes, acting as the support and conflict coefficients, respectively.

In our implementation, we set $\alpha = 0.6$ and $\beta = 0.2$, assigning greater weight to confirmation than to denial. This reflects a fundamental diagnostic principle: the presence of a highly specific symptom is typically stronger positive evidence for a disease than its absence is negative evidence, as symptoms may be unobserved by the patient, or simply delayed in their onset. Previously counted evidence is excluded to avoid double-counting. Since the adjustment factor remains bounded

around the initial *composite score*, the follow-up responses refine rather than replace the initial ranking. The candidates are re-ranked and the most influential answer is identified.

When the final diagnosis is determined, the language model returns to explain. It receives the computed scores and states, for each disease, why it rose or fell in the ranking. Each disease’s confidence is determined by the score margins instead of being asserted: high for a clear leader, cautious when the margin is narrow. The output is therefore a re-ranked differential in which every disease carries a calibrated confidence and an transparent, evidence-based explanation.

IV. EXPERIMENTS AND RESULTS

This section evaluates MedMAS-DDx based on its diagnostic trajectory -its ability to acquire missing information through follow-up questions and iteratively refine the differential diagnosis - rather than as a static classifier.

The three active agents (`symptom_extractor`, `followup_interviewer`, `diagnosis_refiner`) use an LLM via LiteLLM, which can be either a remote API or a self-hosted model deployed through vLLM. All scoring, ranking, and refinement processes run on-premise, independently of the LLM hosting location. Section IV-C traces one evaluation case end to end to illustrate the mechanism concretely.

A. Evaluation of framework and experimental setup

The evaluation of the framework is anchored on the Q4DX benchmark [11], using the `PatientCases.csv` dataset (99 cases, 29 different diagnoses). Performance is evaluated at two levels of symptom completeness, 80% and 50% of the canonical symptoms disclosed in the initial narrative; the fully complete (100%) variant is excluded, since it leaves nothing for follow-up questions to recover. This yields $N = 198$ evaluations (99 per variant).

Follow-up questions are answered by a deterministic simulated patient agent with access to the full symptom profile (case 100%), providing an optimistic estimate of interactive information recovery rather than real-patient performance. Diagnostic correctness is measured using SapBERT cosine similarity with the reference label (target diagnosis), using a 0.6 threshold to tolerate differences in medical terminology.

B. Diagnostic accuracy and interactive refinement

Symptom mapping was verified before assessing accuracy: 91.3% of 1,243 symptoms were mapped to ontology concepts, with the HPO fallback recovering 15.3% of them (e.g., “passage of gases” $\rightarrow$ *flatulence*). As illustrated in Figure 2, the integration of the HPO fallback mechanism significantly reduces early-stage information loss.

Interactive refinement (Phase 2) improves on the zero-shot differential (Phase 1) overall. Among the 42 correct top-1 predictions obtained in Phase 2, only 6 were exact string matches. The remaining cases (e.g., *Hepatitis* vs. *Hepatitis B*) were correctly recognized only through semantic evaluation,

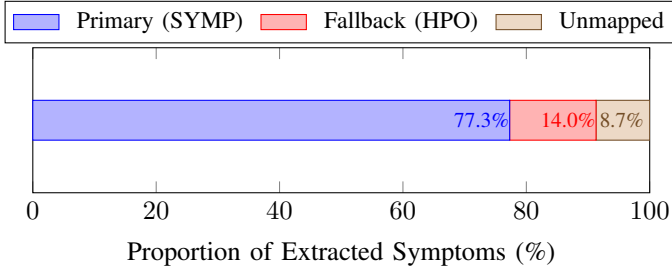


Fig. 2: Distribution of symptom mapping resolution across 1,243 clinical instances

confirming that strict string matching would have underestimated accuracy by almost seven times.

At the aggregate level, interactive refinement (Phase 2) improves on the zero-shot differential (Phase 1) across every metric, including the mean semantic similarity of the top-ranked prediction (Table I).

TABLE I: Global diagnostic accuracy across the reference batch ($N = 198$: 99 evaluations at 80%, 99 at 50%), computed with SapBERT semantic evaluation at the 0.6 threshold. Each metric is reported for Phase 1 (initial differential) and Phase 2 (after follow-up).

Metric	Phase 1	Phase 2
Top-1 accuracy	18.7% (37/198)	21.2% (42/198)
Top-3 accuracy	25.8% (51/198)	27.3% (54/198)
Mean similarity (top-1)	0.449	0.466
Mean similarity (top-3)	0.523	0.537

Table II presents an analysis of the mean semantic similarity per disclosure variant, identifying an important compensatory mechanism to our multi-agent architecture. By engaging in targeted follow-up questions, the system successfully recovers diagnostic confidence, matching the zero-shot baseline of cases with significantly richer initial narratives. This finding demonstrates that the interactive dialogue actively bridges the clinical information gap, effectively compensating for a 30% loss in initial symptom disclosure.

TABLE II: Mean semantic similarity across the reference batch ($N = 198$: 99 evaluations at 80%, 99 at 50%), computed with SapBERT. Each metric is reported for Phase 1 (initial differential) and Phase 2 (after follow-up).

Completeness Variant	Phase 1	Phase 2
80% Completeness	0.4589	0.4792
50% Completeness	0.4394	0.4527

Figure 3 summarizes the results and shows, that refinement is not uniformly advantageous: performance improves from phase 1 to phase 2 in all cases, except for a slight decrease in Top-3 accuracy at 50% completeness. We use the Zero-shot Diagnostic Accuracy (ZDA) metric from the benchmark [11]

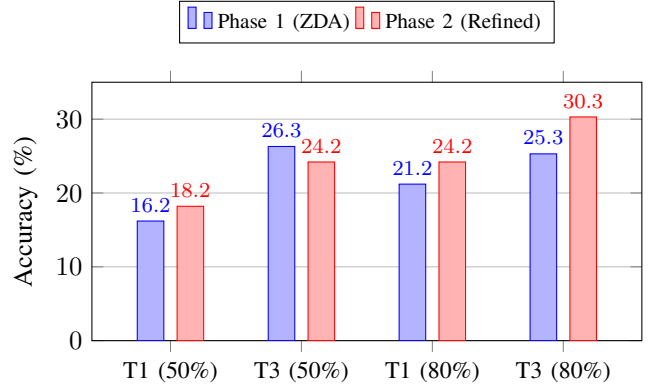


Fig. 3: Top-1 and Top-3 accuracy comparison between the initial Phase 1 and the interactively refined Phase 2.

for phase 1. This asymmetry, together with Case #96 below, is presented in the Discussion.

C. Qualitative case study: Recovering the diagnosis of Hepatitis A (Case #96)

To illustrate how a single follow-up answer can reorder a competitive differential, we trace Case #96 (*Hepatitis A*, 80% completeness) end to end. Its canonical profile includes ten symptoms; the 80% narrative omits *yellowish skin* and *yellowing of eyes* - i.e., jaundice, the most discriminative sign for hepatitis. The clinical narrative given to the pipeline at this completeness level reads as follows:

A 32-year-old male presented to the clinic with troubling symptoms, including persistent joint pain that has worsened over the past two weeks. He reported episodes of vomiting, nausea, and a noticeable loss of appetite. Moreover, he has been suffering from abdominal pain, accompanied by diarrhea. He also noted a mild fever and muscle pain that has hindered his daily activities, which has led to increased concern regarding his health status.

a) *Phase 1a*: The `symptom_extractor` maps all eight symptoms present in the narrative to SYMP concepts (8/8); *jaundice* is simply absent from the text, not lost in mapping.

b) *Phase 1b*: Without the "*jaundice*" symptom, "*gastrointestinal anthrax*" and "*oropharyngeal anthrax*" lead with a *composite score* of 0.922 (Eq. (4)), driven by the high *specificity weight* of "*severe diarrhea*" (Eq. (1)). Four "*hepatitis*" subtypes (A, B, C, E) tie just below at 0.861. The top prediction's SapBERT similarity to "*Hepatitis A*" is only 0.340 - an incorrect Top-1. Although the correct label is tied for third place, the prediction is still considered correct under the Top-3 metric.

c) *Phase 1c*: The `followup_interviewer` asks eleven targeted Yes/No questions (about ascites, acholic stool, dark urine, jaundice, liver inflammation, hematemesis, hematochezia, dyspnea, backache, headache, cough); the simulated patient confirms only "*jaundice*" and denies the others.

d) *Phase 2*: Confirming "jaundice" (Eq. (5)) raises *Hepatitis A, B, and C* to 0.872 despite three denials; "Hepatitis E" gains less, since an added denial on *dark urine* drops it to 0.838. Both "anthrax" candidates receive no support and four denials, falling to 0.781, below all four "hepatitis" subtypes.

e) *Outcome*: "Hepatitis A" becomes the top prediction, with similarity rising from 0.340 to a perfect 1.0 - the largest jump in the entire batch. A single confirmed symptom, combined with tie-aware scoring at the group level, was enough to resolve the ambiguity among six equally scored Phase 1 candidates, which remained unresolved from the initial narrative alone.

To demonstrate the explainability of the final output, Listing IV-C0e shows a subset of the structured reasoning trace generated by the system for the top candidate, *Hepatitis A*.

Listing 1: Agent-generated reasoning trace for the refined diagnosis of Hepatitis A

```

1 "rank": 1,
2 "disease": "hepatitis A",
3 "baseline_score": 0.861,
4 "followup_support": 0.095,
5 "followup_conflict": 0.083,
6 "final_score": 0.872,
7 "updated_confidence": "HIGH",
8 "reasoning": "Hepatitis A initially had a
   score of 0.861. Follow-up questions
   provided support by confirming the
   presence of jaundice (weight +0.733),
   but conflicted by denying ascites (
   weight -0.467), acholic stool (weight
   -0.733), and liver inflammation (
   weight -0.733). The net effect was an
   increase in the final score to 0.872,
   with updated confidence rated as HIGH
   ."

```

D. Comparative analysis against the MedQDx baseline

Figure 4 compares MedMAS-DDx against the MedQDx baseline (GPT-4.1) across both accuracy metrics and completeness levels. MedQDx achieves higher zero-shot Top-1 accuracy (44% at 80%; 37% at 50%), reflecting a different task formulation rather than better reasoning: as an open-vocabulary classifier, it is never penalized for missing labels at the benchmark's exact granularity, whereas MedMAS-DDx grounds each candidate in an explicit knowledge graph and refuses to generate a diagnosis with no ontology correspondent. This constraint enables greater explainability and traceability, as each prediction can be linked to structured evidence and a transparent reasoning process.

Figure 4 shows the same comparison at Top-3, the appropriate granularity for a differential diagnosis: the gap narrows to 30.3% vs. 44% at 80% completeness, but less so at 50% (24.2% vs. 37%), where Top-3 also falls below its own Phase 1 value - the same pattern discussed above.

The two systems also differ sharply in question budget: 2.4 for MedQDx versus roughly ten for MedMAS-DDx, a direct consequence of confirming specific ontology-mapped symptoms rather than a single free-form hypothesis. Whether

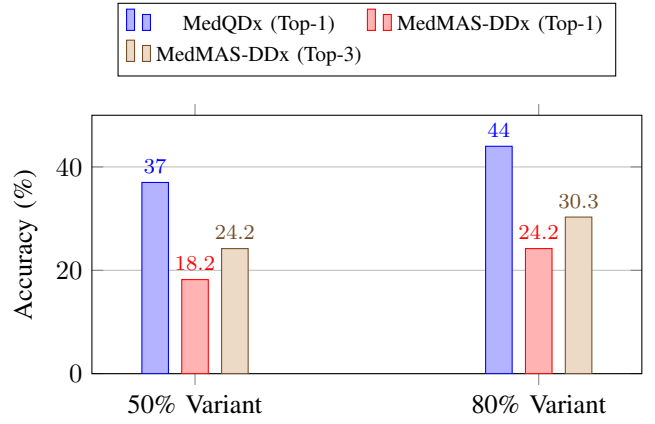


Fig. 4: Performance comparison between MedQDx (unstructured single-label classifier) and MedMAS-DDx (interpretable differential).

this cost is acceptable depends on the deployment context, discussed further in the Discussion.

V. DISCUSSIONS

In this section, we discuss the behavioral characteristics and the possible limitations of the proposed system.

An important characteristic of our approach is the use of workflow orchestration to coordinate specialized agents through deterministic control logic instead of unconstrained agent-to-agent communication. This architecture provides predictable execution and allows every intermediate decision to be validated before subsequent processing. Consequently, the system remains suitable for safety-critical clinical applications where reproducibility and auditability are essential.

The system leverages a multi-turn questioning mechanism that progressively enriches the clinical picture through iterative information gathering. When the initial input is sparse or incomplete, the interaction naturally shifts toward targeted follow-up questions that reduce uncertainty by collecting discriminative clinical evidence. Each turn refines the differential diagnosis by incorporating newly acquired responses into the ranking process, enabling a gradual convergence toward more accurate conclusions. This iterative information-gathering process highlights the strength of structured dialogue in improving diagnostic precision while maintaining a transparent and traceable reasoning flow.

Overall, the contribution of interactive refinement is real but modest, and at 50% completeness it is not even uniformly positive: Top-1 accuracy improves at both completeness levels (Fig. 3), but Top-3 accuracy improves only at 80% completeness (25.3% → 30.3%) and decreases slightly at 50% (26.3% → 24.2%), a pattern that also makes the narrowing of the gap with the MedQDx baseline uneven across variants.

The case study in Section IV-C suggests a different aspect. Case #96 does not move the aggregate accuracy by a fraction of a point; it moves one case's Top-1 similarity from 0.340 to a perfect 1.0, by resolving the ambiguity among six equally

scored Phase 1 candidates with a single targeted confirmation. When ambiguity exists in the initial ranking, refinement resolves it decisively. However, when one diagnosis already dominates or no initial candidate is sufficiently close to the reference label, refinement has no actionable uncertainty to resolve, and its measured contribution is zero by construction.

The Top-3 drop at 50% completeness arises because incomplete symptom information allows weak candidates to enter the Top-3. Follow-up evidence can then legitimately replace these incidental candidates through the score update (Eq. (4)) with a better-supported candidate, occasionally reducing Top-3 accuracy. Aggregating cases where refinement can resolve ambiguity with those where it does not inherently understates its per-case impact; but case-level trajectories provide a more faithful assessment of the follow-up loop's contribution.

Despite the advantages, several limitations should be acknowledged. First, the quality of the differential diagnosis depends on the completeness and accuracy of the underlying biomedical ontology and its symptom-disease associations. Missing, incomplete, or outdated knowledge cannot be incorporated into the reasoning process.

Other key limitation in medical diagnosis tasks is the mismatch between ontology terminology and dataset terminology, which arises from the gap between formal, structured medical knowledge (ground truth) and the context-dependent, heterogeneous language used in clinical data as it is recorded in real-world practice. This gap limits the alignment of patient data with ontology concepts, affecting tasks such as symptoms-disease mapping or information extraction. This limitation is partially mitigated by the extensive validation and widespread adoption of biomedical ontologies in research and several clinical systems, [20], [21], [22] that explicitly leverage the DOID or the HPO and which provide a robust semantic foundation for clinical reasoning.

The limitation of the current knowledge base is exposed by two reference labels, *Jaundice* and *Vertigo*: the Q4Dx benchmark treats both as diagnoses, but SYMP classifies both as symptoms. Since each candidate MedMAS-DDx proposes must correspond to an explicit disease concept in the knowledge graph, the system cannot output either label as a diagnosis, regardless of how well the underlying clinical picture is understood - a direct consequence of ontology coverage, not of the reasoning pipeline.

For *Jaundice*, the limitation is absolute: it is never proposed as a candidate disease across the six evaluations (three cases, both completeness levels). The system instead retrieves ontology-supported diseases associated whose profile includes *jaundice* as a symptom (e.g., *Hepatitis E*, *D*, *hepatic tuberculosis*), and in four of the six evaluations asks about *jaundice* as follow-up evidence, but accuracy remains zero until the ontology is extended with a matching concept, regardless of reasoning quality.

Another limitation is the restricted interaction model. Follow-up questions are limited to "yes/no/unsure" responses to support deterministic scoring. Although this improves reproducibility and simplifies reasoning, it cannot fully capture

the clinical information such as symptom severity, temporal progression, or uncertainty that may influence diagnostic decisions.

VI. CONCLUSION AND FUTURE WORK

Overall, the proposed methodology forms a directed pipeline with a single feedback loop: free text is transformed into structured symptoms, symptoms generate candidate diseases, and candidates are consolidated into an initial ranking that either terminates or triggers targeted questioning. Follow-up answers are then reintegrated into the scoring stage. Throughout the pipeline, each component has a distinct role: language models perform interpretation and generation, the external knowledge source provides symptom-disease associations, and deterministic code handles ranking, evaluation, and control flow. A sparse initial description simply produces a weaker initial ranking and moves more of the work onto the questioning stage, so the amount of questioning adapts to how uncertain the starting point is. This consistency, more than any single component, keeps the process transparent at every stage.

Despite the promising results, several opportunities for future development remain. The system should be validated on large-scale real-world dataset such as MIMIC-IV [23], [24] or DDXPlus [25], whose clinical notes, laboratory results, demographics, and diverse patient cases better reflect real-world complexity. Such evaluation would address the limitations of the current benchmark, which is based on a relatively limited dataset.

Future work will focus on extending ontology coverage, incorporating richer patient interactions while preserving deterministic reasoning, and performing prospective clinical evaluation. In addition, adaptive weighting strategies and the integration of probabilistic evidence could further improve the accuracy of the ranking without reducing the transparency of the decision-making process.

Acknowledgements. We gratefully acknowledge Georgiana Copil (Adobe Romania) for her guidance and coordination throughout the development of this work. Roxana Szomiu is supported by a grant provided by Adobe Romania.

REFERENCES

- [1] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, "What disease does this patient have? a large-scale open domain question answering dataset from medical exams," *Applied Sciences*, vol. 11, no. 14, 2021. [Online]. Available: <https://www.mdpi.com/2076-3417/11/14/6421>
- [2] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering," 2022. [Online]. Available: <https://arxiv.org/abs/2203.14371>
- [3] D. McDuff, M. Schaekermann, T. Tu, A. Palepu, A. Wang, J. Garrison, K. Singhal, Y. Sharma, S. Azizi, K. Kulkarni, L. Hou, Y. Cheng, Y. Liu, S. S. Mahdavi, S. Prakash, A. Pathak, C. Semturs, S. Patel, D. R. Webster, E. Dominowska, J. Gottweis, J. Barral, K. Chou, G. S. Corrado, Y. Matias, J. Sunshine, A. Karthikesalingam, and V. Natarajan, "Towards accurate differential diagnosis with large language models," 2023. [Online]. Available: <https://arxiv.org/abs/2312.00164>
- [4] L. M. Schriml, C. Arze, S. Nadendla, A. Ganapathy, V. Felix, A. Mahurkar, K. Phillippy, A. Gussman, S. Angiuoli, E. Ghedin, O. White, and N. Hall, "GeMInA, genomic metadata for infectious agents, a geospatial surveillance pathogen database," *Nucleic Acids Res.*, vol. 38, no. Database issue, pp. D754–64, Jan. 2010.

- [5] B. JA, J. CMS, S. MA, O. D, N. L, F. V, B. SM, G. C, L. R, B. K, K. R, B. C, M. R, A. N, B. K, B. E, C. L, D. G, H. H, D. S, S. P, G. J, F. M, G. P, Z. S, C. R, E. S, W. A, U. Y, L. E, C. M, D. T, and S. LM, "The do-kb knowledgebase 2026 update: expanding programmatic and language access," *Nucleic Acids Research*, vol. 54, no. D1, pp. D1425–D1436, 2026. [Online]. Available: <https://doi.org/10.1093/nar/gkaf1213>
- [6] L. M. Schriml, J. B. Munro, M. Schor, D. Olley, C. McCracken, V. Felix, J. A. Baron, R. Jackson, S. M. Bello, C. Bearer, R. Lichenstein, K. Bisordi, N. C. Dialo, M. Giglio, and C. Greene, "The human disease ontology 2022 update," *Nucleic Acids Res.*, vol. 50, no. D1, pp. D1255–D1261, Jan. 2022.
- [7] P. N. Robinson, S. Köhler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos, "The human phenotype ontology: a tool for annotating and analyzing human hereditary disease," *Am. J. Hum. Genet.*, vol. 83, no. 5, pp. 610–615, Nov. 2008.
- [8] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, M. Amin, L. Hou, K. Clark, S. R. Pfohl, H. Cole-Lewis, D. Neal, Q. M. Rashid, M. Schaekermann, A. Wang, D. Dash, J. H. Chen, N. H. Shah, S. Lachgar, P. A. Mansfield, S. Prakash, B. Green, E. Dominowska, B. Agüera Y Arcas, N. Tomašev, Y. Liu, R. Wong, C. Semturs, S. S. Mahdavi, J. K. Barral, D. R. Webster, G. S. Corrado, Y. Matias, S. Azizi, A. Karthikesalingam, and V. Natarajan, "Toward expert-level medical question answering with large language models," *Nat. Med.*, vol. 31, no. 3, pp. 943–950, Mar. 2025. [Online]. Available: <https://www.nature.com/articles/s41591-024-03423-7>
- [9] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, N. Scharli, A. Chowdhery, P. Mansfield, B. A. y Arcas, D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Semturs, A. Karthikesalingam, and V. Natarajan, "Large language models encode clinical knowledge," 2022. [Online]. Available: <https://arxiv.org/abs/2212.13138>
- [10] G. A. Muntean, A. Marginean, A. Groza, I. Damian, S. A. Roman, M. C. Hapca, A. M. Sere, R. M. Mănoiu, M. V. Muntean, and S. D. Nicoară, "A qualitative evaluation of chatgpt4 and palm2's response to patient's questions regarding age-related macular degeneration," *Diagnostics*, vol. 14, no. 14, 2024. [Online]. Available: <https://www.mdpi.com/2075-4418/14/14/1468>
- [11] M. Werthaim, M. Kimhi, A. Apartsin, and Y. Apterstein, "A benchmark for evaluating diagnostic questioning efficiency of LLMs in patient conversations," *Sci. Rep.*, vol. 16, no. 1, p. 6121, Jan. 2026. [Online]. Available: <https://www.nature.com/articles/s41598-026-37022-y>
- [12] J. Tucker, "SymbiPredict," 2024. [Online]. Available: <https://data.mendeley.com/datasets/dv5z3v2xyd/1>
- [13] Q. Wang, R. Sheng, Y. Li, H. Qu, Y. Sun, and M. Zhu, "Medkgi: Iterative differential diagnosis with medical knowledge graphs and information-guided inquiring," 2026. [Online]. Available: <https://arxiv.org/abs/2512.24181>
- [14] D. F. V., O. Proniak, V.-E. Gruber, and R. Marinescu, "Medpi: Evaluating ai systems in medical patient-facing interactions," 2025. [Online]. Available: <https://arxiv.org/abs/2601.04195>
- [15] J. Fang, R. Chen, X. Yang, J. Yu, J. Xu, A. Vinod, W. Shi, T. Chen, H. Ji, C. Zhai, Y. Ding, and Y. Zhang, "Benchmarking multi-turn medical diagnosis: Hold, lure, and self-correction," 2026. [Online]. Available: <https://arxiv.org/abs/2604.04325>
- [16] L. Gong, W. Fang, T. Yang, D. Tao, C. Guo, P. Wei, B. Xie, J. Guan, Z. Chen, F. Shi, J. Gu, and J. Liu, "Meddialogrubrics: A comprehensive benchmark and evaluation framework for multi-turn medical consultations in large language models," 2026. [Online]. Available: <https://arxiv.org/abs/2601.03023>
- [17] X. Tang, A. Zou, Z. Zhang, Z. Li, Y. Zhao, X. Zhang, A. Cohan, and M. Gerstein, "Medagents: Large language models as collaborators for zero-shot medical reasoning," 2024. [Online]. Available: <https://arxiv.org/abs/2311.10537>
- [18] G. Stinton, J. A. Lieviant, S. Kam, J. Y. Lim, J. C.-Y. Goh, W. K. Lim, G. Baynam, T. Tan, D.-S. Pham, and S. S. Jamuar, "Clinical free text to hpo codes," *Rare*, vol. 1, p. 100007, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2950008723000078>
- [19] W. Zhong, M. Sun, S. Yao, Y. Liu, D. Peng, Y. Liu, K. Yang, H. Gao, H. Yan, W. Hao, Y. Yan, and C. Yin, "Enhancing the accuracy of human phenotype ontology identification: Comparative evaluation of multimodal large language models," *J Med Internet Res*, vol. 27, p. e73233, Jun. 2025. [Online]. Available: <https://www.jmir.org/2025/1/e73233>
- [20] J. Zhang, X. Zhang, K. Sun, X. Yang, C. Dai, and Y. Guo, "Unsupervised annotation of phenotypic abnormalities via semantic latent representations on electronic health records," 2019. [Online]. Available: <https://arxiv.org/abs/1911.03862>
- [21] L. M. Schriml, C. Arze, S. Nadendla, Y.-W. W. Chang, M. Mazaitis, V. Felix, G. Feng, and W. A. Kibbe, "Disease ontology: a backbone for disease semantic integration," *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D940–6, Jan. 2012. [Online]. Available: <https://doi.org/10.1093/nar/gkr972>
- [22] S. Köhler, M. Gargano, N. Matentzoglou, L. C. Carmody, D. Lewis-Smith, N. A. Vasilevsky, D. Danis, G. Balagura, G. Baynam, A. M. Brower, T. J. Callahan, C. G. Chute, J. L. Est, P. D. Galer, S. Ganesan, M. Griesse, M. Haimel, J. Pazmandi, M. Hanauer, N. L. Harris, M. Hartnett, M. Hastreiter, F. Hauck, Y. He, T. Jeske, H. Kearney, G. Kindle, C. Klein, K. Knoflach, R. Krause, D. Lagorce, J. A. McMurry, J. A. Miller, M. Munoz-Torres, R. L. Peters, C. K. Rapp, A. M. Rath, S. A. Rind, A. Rosenberg, M. M. Segal, M. G. Seidel, D. Smedley, T. Talmy, Y. Thomas, S. A. Wiafe, J. Xian, Z. Yüksel, I. Helbig, C. J. Mungall, M. A. Haendel, and P. N. Robinson, "The human phenotype ontology in 2021," *Nucleic Acids Research*, vol. 49, no. D1, pp. D1207–D1217, 01 2021. [Online]. Available: <https://doi.org/10.1093/nar/gkaa1043>
- [23] A. Johnson, L. Bulgarelli, T. Pollard, B. Gow, B. Moody, S. Horng, L. A. Celi, and R. Mark, "MIMIC-IV," *PhysioNet*, Oct. 2024, version 3.1. [Online]. Available: <https://doi.org/10.13026/kpb9-mt58>
- [24] B. Stark, "Disease and symptoms dataset 2023," 2025. [Online]. Available: <https://data.mendeley.com/datasets/2cxcscxydc/1>
- [25] A. F. Tchango, R. Goel, Z. Wen, J. Martel, and J. Ghosn, "Ddxplus: A new dataset for automatic medical diagnosis," 2022. [Online]. Available: <https://arxiv.org/abs/2205.09148>