OntoDiagnose: Ontology-Grounded Reasoning for Medical Question Answering with Language Models

Pratik Jalan[1]* **Mukul Joshi**[1]* **Kshitij Jadhav**[1]
[1]Indian Institute of Technology Bombay, Mumbai, India
Corresponding author: kshitij.jadhav@iitb.ac.in

Abstract

Language models achieve strong performance in medical question answering, yet hallucinations persist even in domain-specialized models, particularly on complex clinical vignettes involving rare presentations, multi-hop reasoning, and ambiguous evidence requiring explicit relational grounding. We introduce OntoDiagnose, a training-free, ontology-grounded, clinician-inspired reasoning pipeline that augments parametric knowledge with structured evidence from SNOMED CT. Given a clinical question, OntoDiagnose extracts salient medical entities, grounds them in the ontology, explores a bounded subgraph, and compresses relevant relations into compact, interpretable evidence chains that guide reasoning. We evaluate OntoDiagnose across USMLE-style and biomedical QA benchmarks using MedGemma, BioMistral, MedPhi, and OpenBioLLM under zero-shot, few-shot, and chain-of-thought settings. OntoDiagnose consistently improves over prompt-only baselines, with an average gain of 9.6 accuracy points on MedQA and the largest relative gains for the lowest-scoring models. These results show the benefit of incorporating ontology-grounded evidence into medical question answering without modifying model parameters.

1 Introduction

Large language models (LLMs) have revolutionized natural language processing, yet their application to medical question answering continues to expose fundamental limitations. In complex clinical scenarios such as USMLE-style vignettes involving multi-hop reasoning across symptoms, signs, and diagnoses, LLMs frequently hallucinate, generating plausible but medically incorrect responses (Singhal et al., 2023; Farquhar et al., 2024; Thirunavukarasu et al., 2023). These failures stem from the inherently probabilistic nature of LLMs: trained to model surface-level statistical regularities, they often exhibit brittle recall, unsupported inference chains, and sensitivity to prompt phrasing (Kim et al., 2025; WU et al., 2024). Even domain-specialized medical LLMs remain vulnerable to these issues, particularly for rare disease presentations, ambiguous clinical evidence, or reasoning scenarios that deviate from training distributions (Liévin et al., 2024; Nori et al., 2023). While fine-tuning or scaling model parameters can partially alleviate these limitations, such approaches are resource-intensive and do not directly address the core challenge of grounded, relational reasoning in high-stakes clinical settings. Clinical decision making requires explicit navigation of structured relationships between symptoms, findings, investigations, and diagnoses; such relationships are not reliably encoded in parametric model weights alone. This gap motivates the integration of exter-

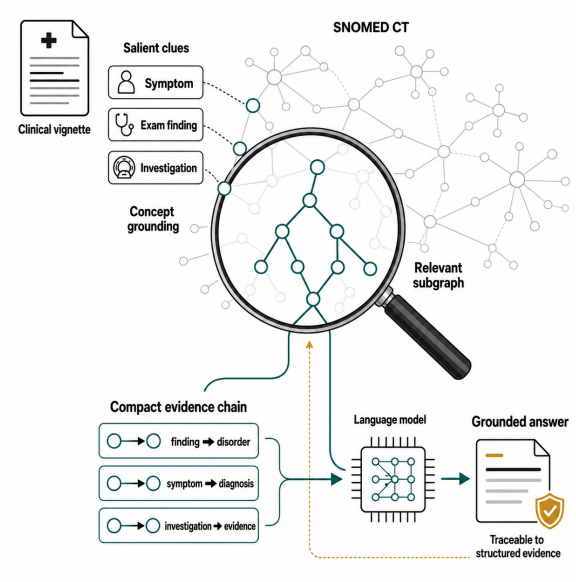


Figure 1: Overview of OntoDiagnose. Vignette clues are grounded in SNOMED CT and the surrounding subgraph compressed into evidence chains.

*Pratik Jalan and Mukul Joshi contributed equally to this work and share first authorship.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 37819–37843
October 24-29, 2026 ©2026 Association for Computational Linguistics

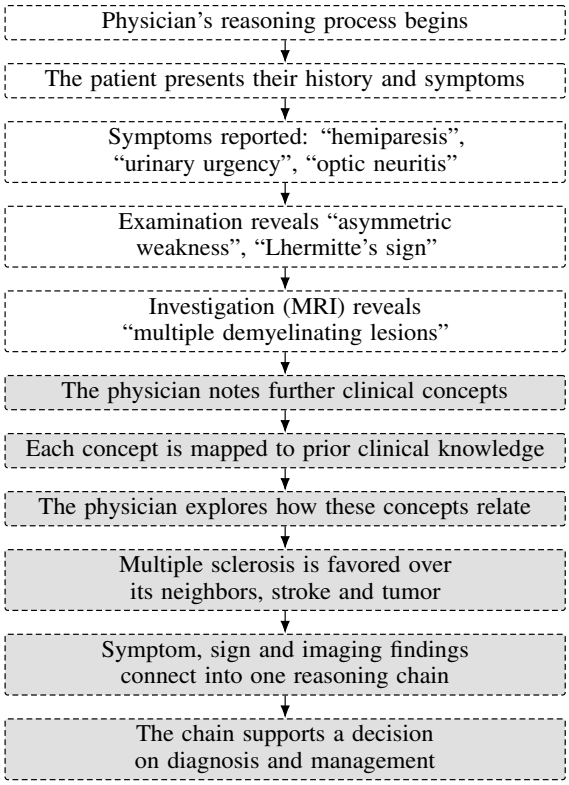


Figure 2: The diagnostic workflow that motivates OntoDiagnose. Findings are gathered (unshaded), then connected into one diagnostic chain (shaded).

nal structured biomedical knowledge to constrain and guide model reasoning in a clinically meaningful way. Figure 1 sketches how OntoDiagnose does so: the clues a vignette offers are grounded in SNOMED CT, the surrounding subgraph is compressed into a handful of evidence chains, and those chains are what the language model reads before it commits to an answer, so the support for an answer stays visible alongside it.

Clinical motivation: Consider the following illustrative case. A 27-year-old woman presents to the emergency department with progressive left-sided hemiparesis, urinary urgency, and a prior episode of optic neuritis. On examination, she demonstrates asymmetric weakness and a positive Lhermitte's sign, while magnetic resonance imaging reveals multiple demyelinating lesions. Figure 2 traces the reasoning such a case invites. The clinician separates salient symptoms from the narrative ("hemiparesis", "urinary urgency", "optic neuritis"), adds examination and imaging findings ("asymmetric weakness", "Lhermitte's sign", "multiple demyelinating lesions"), grounds each observation in prior medical knowledge, explores how those concepts relate, and connects them into a chain favoring multiple sclerosis over alternatives such as stroke or neoplasm. This process of

extract → ground → explore → connect

is what OntoDiagnose automates end to end.

Current gap: Despite advances in prompt engineering and reasoning-style prompting, direct zero-shot or few-shot application of LLMs to medical question answering remains unreliable. Language-only approaches struggle to explicitly represent and traverse structured biomedical relations, while purely symbolic systems such as biomedical ontologies encode rich, curated knowledge but lack the ability to interpret natural language clinical narratives. Bridging this divide between unstructured clinical text and structured medical knowledge remains an open challenge.

Our approach: We propose ONTODIAGNOSE, a clinician-inspired, ontology-grounded reasoning pipeline (Figure 3) designed to support medical question answering by explicitly injecting structured biomedical knowledge into LLM inference. OntoDiagnose follows the stages below:

1. **Entity recognition and grounding:** Clinical concepts are extracted from the vignette using domain-aligned encoders and mapped to standardized ontology nodes (e.g., SNOMED CT), forming a seed set of medically relevant entities.
2. **Ontology exploration:** The system expands these seeds within a bounded neighborhood of the ontology graph to surface diagnostically relevant findings, investigations, and disorders.
3. **Graph summarization:** The explored subgraph is compressed into a compact connector structure analogous to a minimal spanning tree, preserving salient multi-hop relations while suppressing noise.
4. **Evidence serialization:** The resulting paths are rendered as short, interpretable evidence statements (e.g., "Optic neuritis is a common presenting feature of multiple sclerosis").
5. **LLM reasoning:** These ontology-derived evidence statements are injected into the model prompt alongside the question stem and answer options, guiding the language model toward clinically grounded inference.

This workflow is motivated by a simplified diagnostic process and combines structured knowledge navigation with language-model answer selection, supported by inspectable evidence.

Our contributions are threefold:

1. **A reusable knowledge-grounding interface for LMs:** We introduce ONTODIAGNOSE, a training-free pipeline that converts unstructured clinical text into compact, structured evidence via three interacting design choices: (a) clinically constrained entity grounding that restricts seed selection to domain-relevant semantic types; (b) bounded ontology exploration with degree-penalized traversal, replacing unconstrained multi-hop walks; and (c) metric-closure MST compression that distills explored subgraphs into minimal tree-shaped connectors. This pattern (ground $\rightarrow$ explore $\rightarrow$ compress $\rightarrow$ serialize) is not specific to SNOMED CT and could in principle be instantiated over other structured knowledge graphs, though we evaluate it only in the clinical setting.
2. **Interpretable evidence chains with clinical evaluation:** Every evidence statement corresponds to a named path through SNOMED CT, so a reader can check a chain against the ontology instead of against the model. Two physicians rated 78.3% of these chains as fully supporting the selected answer, with Cohen's $\kappa = 0.81$ between them, and controlled ablations place the gains with structured ontology grounding rather than with token-length effects, prompt engineering, or surface-form overlap.
3. **Systematic evaluation showing larger benefit for lower-scoring LMs:** We evaluate OntoDiagnose on five QA benchmarks with four medical LMs under zero-shot, few-shot, and CoT, and add two larger systems under zero-shot prompting to test whether the benefit survives at scale. Gains scale inversely with baseline accuracy: across all six, the lowest-scoring model improves by 33% relative to its own zero-shot score and the highest-scoring by 4%. Explicit relational structure therefore appears to substitute for parametric knowledge where that knowledge is thinnest, a finding that motivates further evaluation in resource-constrained settings.

2 Related Work

Prior approaches: Medical LLMs such as Med-PaLM (Singhal et al., 2022) encode substantial clinical knowledge yet remain prone to hallucinations and multi-hop reasoning failures (Jin et al., 2020; Liévin et al., 2024; Tang et al., 2023). RAG conditions generation on retrieved passages (Lewis et al., 2021; Banerjee et al., 2026; Jalan et al., 2026), which supplies external evidence but does not by itself guarantee factual multi-hop clinical reasoning, and document-centric retrieval cannot explicitly traverse structured clinical relations. Biomedical knowledge graphs (SNOMED CT, UMLS, PrimeKG) (El-Sappagh et al., 2018; Chandak et al., 2022) encode curated relations supporting multi-hop inference (Rao et al., 2022), and KG-augmented LLM methods (Soman et al., 2024; Edge et al., 2024) integrate this structure into prompting pipelines. QA-GNN (Yasunaga et al., 2021), GreaseLM (Zhang et al., 2022), and G-Retriever (He et al., 2024) fuse LM representations with graph structure, while concurrent works (ReGraM (Lee et al., 2026), condition-gated reasoning (Parekh et al., 2026), multi-round agentic retrieval (Wu et al., 2026), DEEPMED (Wang et al., 2026)) explore region-bounded, condition-gated, and agentic strategies. However, most pipelines employ generic traversal without clinically constrained, compact evidence structures. Extended discussion of all related methods, entity linking comparisons (SapBERT (Liu et al., 2021) vs. SNOBERT (Kulyabin et al., 2024) and the multiple-choice formulation of (Lin et al., 2024)), and concurrent KG reasoning approaches is provided in Appendix B.

Positioning OntoDiagnose: Table 1 situates our work. Prior medical LLMs rely on parametric knowledge; RAG adds document evidence but lacks explicit multi-hop reasoning; KG-QA systems enable structured inference but cannot interpret free-text narratives. OntoDiagnose combines LLMs with ontology-bounded, clinician-aligned graph exploration and MST-based evidence compression, following the structure of a simplified diagnostic workflow while remaining training-free and compatible with resource-efficient models.

3 Methods

This section specifies ONTODIAGNOSE, a clinician-inspired, ontology-guided pipeline for

Method	Model		Knowledge Access		Reasoning Support		
	Domain	LM	Retrieval	KG	Multi-hop Reasoning	Clinical Grounding	Compressed Evidence
Parametric language models							
Med-PaLM (Singhal et al., 2022)	Medical	✓	✗	✗	△	✗	✗
Prompted general-purpose LMs (Nori et al., 2023; Liévin et al., 2024)	General	✓	✗	✗	△	✗	✗
Retrieval-augmented methods							
RAG (Lewis et al., 2021)	General	✓	✓	✗	△	✗	✗
Medical RAG (Xiong et al., 2024; Zakka et al., 2023)	Medical	✓	✓	✗	△	✗	✗
Graph-based methods							
Biomedical KG-QA (Rao et al., 2022)	Medical	✗	✓	✓	✓	✓	✗
KG / Graph RAG (Edge et al., 2024; Soman et al., 2024)	General / Medical	✓	✓	✓	✓	△	✗
OntoDiagnose (ours)	**Medical**	✓	✓	✓	✓	✓	✓

Table 1: Comparison of representative approaches to medical question answering. ✓ denotes full support, △ partial support, and ✗ absence of a property. LM denotes language model and KG denotes knowledge graph. Compressed evidence indicates whether retrieved graph structure is distilled into a compact connector before prompting.

medical question answering. Training-free denotes no parameter updates; inference-time costs are reported in Appendix D.2.

3.1 Overview of the OntoDiagnose framework

Figure 3 summarizes the end-to-end workflow of OntoDiagnose. The process begins with a free-text clinical vignette (**Step 1**), which is parsed to extract salient clinical mentions such as symptoms, examination findings, and investigations (**Steps 2–3**). These mentions are grounded to standardized SNOMED CT concepts (**Steps 4–5**), transforming unstructured clinical text into a graph-based representation. The grounded concepts are then expanded within a bounded neighborhood of the ontology (**Steps 6–7**) to produce a structured subgraph that captures diagnostically relevant relations. This subgraph is compressed into a compact connector structure and serialized into concise evidence statements (**Steps 8–10**), which summarize clinically meaningful multi-hop relations. Finally, the evidence statements are combined with the original question and answer options (**Step 11**) and provided as contextual input to a language model (**Step 12**), which integrates this structured evidence with its parametric knowledge to select the most plausible answer (**Step 13**). A detailed running example illustrating each pipeline stage with SNOMED CT paths is provided in Appendix H.

3.2 Ontology: SNOMED CT International

We use SNOMED CT, a large-scale, multilingual clinical ontology maintained by SNOMED International (El-Sappagh et al., 2018), as the biomedical knowledge backbone of OntoDiagnose (El-Sappagh et al., 2018). SNOMED CT consists of standardized concepts representing diseases, findings, procedures, and anatomy, organized via poly-hierarchical is-a relations and additional semantic attributes (e.g., finding site, associated morphology) (El-Sappagh et al., 2018). These curated relations enable structured, multi-hop navigation across clinically meaningful concepts.

3.3 Problem setup and notation

We consider a dataset of clinical cases paired with multiple-choice questions:

$$\mathcal{D} = \{((C_i, Q_i), A_i, Y_i)\}_{i=1}^{N},$$

where C_i denotes a free-text clinical vignette (history, vitals, examinations, imaging), Q_i the question stem, $A_i = \{a_{i1}, \ldots, a_{iM}\}$ the candidate answers, and Y_i the correct label index. Structured medical knowledge is represented as a directed, labeled graph

$$\mathcal{G} = (V, E),$$

where each node $v \in V$ corresponds to a canonical medical concept (e.g., disease, finding, procedure, anatomy) and each edge $e \in E$ encodes a semantic relation (e.g., "has-finding", "associated-with", "has-site"). Given a clinical case, OntoDiagnose extracts relevant concepts, grounds them in ontology nodes, explores their local neighborhood, and distills the resulting subgraph into compact evidence chains that support answer selection by a medical

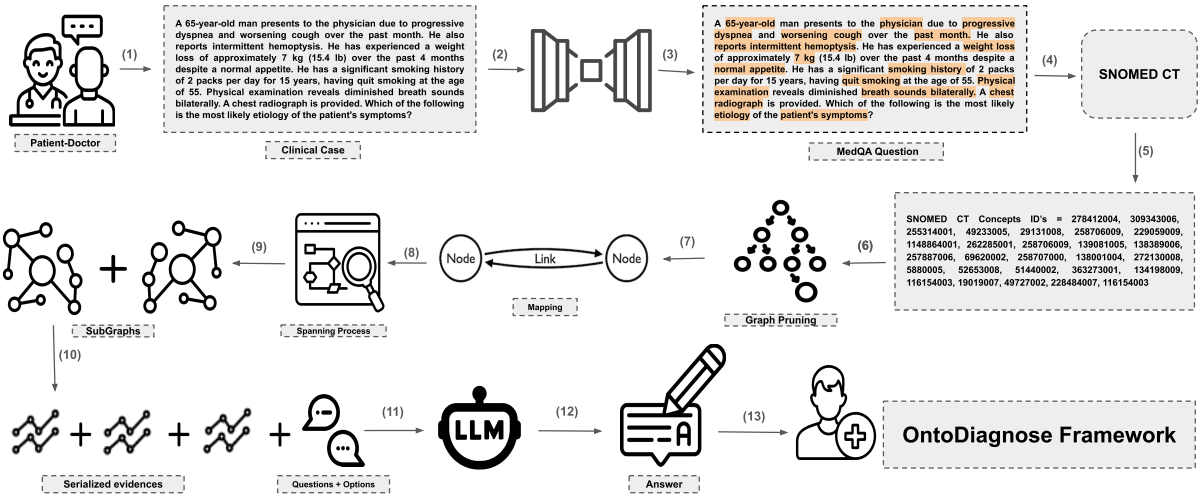


Figure 3: The OntoDiagnose pipeline, from the input vignette to the selected answer. A vignette is parsed into salient medical entities, which are grounded to SNOMED CT concepts; a bounded neighborhood is explored and compressed into multi-hop evidence chains, and the serialized chains enter the prompt alongside the question and options. All steps described in Section 3.1.

language model. **Our pipeline is a composition of the following modules:**

$$\text{Case} \xrightarrow{f_{\text{parse}}} T \xrightarrow{f_{\text{NER}}} E \xrightarrow{f_{\text{link}}} S \xrightarrow{\text{explore}} G_S$$
$$G_S \xrightarrow{\text{summarize}} T_s \xrightarrow{\text{serialize}} P \xrightarrow{\pi_\theta} \hat{A}.$$

where S denotes the set of grounded ontology seed nodes, G_S the explored ontology subgraph, T_s its compact summarized structure, and P the serialized evidence-infused prompt provided to the language model π_θ.

3.4 Ontology-Grounded Evidence Construction

OntoDiagnose converts unstructured clinical vignettes into compact, interpretable evidence structures through a multi-stage ontology-grounded pipeline:

Entity extraction and grounding: Salient medical mentions are extracted using a hybrid strategy combining high-precision lexical matching (for canonical terms) with neural biomedical encoders (SciSpaCy for segmentation, SapBERT for contextual embeddings) (Neumann et al., 2019; Liu et al., 2021). Extracted mentions are grounded to canonical SNOMED CT concepts via embedding-based approximate nearest-neighbor search (FAISS) (Johnson et al., 2019) and lightweight contextual disambiguation using semantic type constraints and sentence-level reranking. We adopt SapBERT for its strong published performance on biomedical entity normalization. On our held-out set it reaches 84.7% top-1 accuracy, against 86.3% for the SNOMED-specific SNOBERT at 18 ms per mention versus 14 ms, so SapBERT offers a favorable rather than a uniformly better accuracy-latency trade-off (Appendix I). Entity linking remains the largest single source of downstream error in our analysis (31% of failures), and we return to it in the Limitations.

Bounded ontology exploration: Starting from the grounded seed nodes, the system performs bounded breadth-first traversal (radius r=3, max path length $L_{\max}$=5) to identify short, clinically meaningful multi-hop paths connecting observed findings to diagnostically relevant concepts. We traverse seven SNOMED CT relation types (is-a, finding-site, associated-morphology, causative-agent, has-definitional-manifestation, due-to, occurrence), which together account for 97.3% of the edges in the explored neighborhoods. By default the traversal ignores edge direction. This is a deliberate trade-off rather than an oversight: undirected traversal yields an evidence chain for 94.1% of questions against 78.6% under strict direction-preserving traversal, because many diagnostic paths ascend the is-a hierarchy before descending to a related disorder. The cost is that a minority of paths traverse a non-hierarchical relation against its authored direction, and we quantify the resulting accuracy and coverage of a direction-preserving hybrid in Appendix G.7. High-degree hub nodes are penalized by inverse-degree

Model	ZS	FS	CoT	Ours	Domain	Gain
MedGemma (4B)	50.92	51.76	52.12	**61.19**	Medical	+10.27
OpenBioLLM (8B)	45.01	46.56	46.95	53.79	Medical	+8.78
MedPhi (3B)	43.96	46.12	46.91	51.97	Medical	+8.01
BioMistral (7B)	35.18	38.20	41.18	46.69	Medical	**+11.51**
Average	**43.77**	**45.66**	**46.79**	**53.41**	–	**+9.64**
MedGemma (27B)	62.41	–	–	68.39	Medical	+5.98
Gemini 2.5 Flash	71.53	–	–	74.29	General	+2.76

Table 2: Accuracy (%) on MedQA across prompting regimes. ZS, FS, and CoT denote zero-shot, few-shot, and chain-of-thought prompting, respectively. Bold marks the best result or gain among the four primary models, and the shaded row reports their average.

weighting above a degree threshold of 500 (Appendix G.2).

Graph summarization: The explored subgraph is compressed into a minimal, tree-shaped connector via metric-closure MST approximation (Kou et al., 1981; Mehlhorn, 1988), preserving salient multi-hop relations while suppressing irrelevant branches. We adopt uniform edge weights (w_e=1) as a transparent, training-free baseline; ablations show that relation-type weighting yields only modest gains (+0.48 pp), so the bounded traversal and MST compression account for most of the benefit (Appendix G.7). We use metric-closure MST as a transparent, training-free compression strategy, not as a claim that it is an optimal graph-selection objective; learned path selection or reranking is a natural extension, but would require supervision and would forgo the training-free property.

Evidence serialization: From the compact tree, a small number of evidence chains are rendered as concise, templated natural-language statements (e.g., "Steatorrhea is due to Pancreatic exocrine insufficiency, which is due to Cystic fibrosis"). These chains verbalize ontology-encoded relationships and serve as structured external evidence for the downstream language model. Full algorithmic details are provided in Appendix C and pseudocode in Algorithm 1.

3.5 Prompt construction and LLM answer selection

The final prediction is produced by a language model conditioned on the original clinical question and a small set of serialized ontology-derived evidence chains. Rather than eliciting unconstrained free-form reasoning, OntoDiagnose injects structured, externally generated evidence into the prompt to guide inference in a controlled manner. The prompt combines the clinical context, serialized evidence chains, the question stem, and answer options (the full template is provided in Appendix A.2). We evaluate four inference settings: zero-shot, where the model receives only the clinical context and question; few-shot, where it is provided with three example question-answer pairs; chain-of-thought, which asks the model to reason step by step without external knowledge; and OntoDiagnose, which augments the input with ontology-derived evidence chains to enable evidence-guided inference (Wei et al., 2022; Lewis et al., 2021).

Answer selection: For multiple-choice questions, we score candidate answers using the model's log-probabilities over explicit option tokens when available, selecting

$$\hat{a} = \arg\max_{a \in \mathcal{A}} \log \pi_\theta(a \mid q, \mathcal{C}),$$

where $\mathcal{C}$ denotes the combined clinical context and injected evidence.

4 Experiments and Results

4.1 Experimental Setup

We evaluate ONTODIAGNOSE under controlled prompting settings across four openly available medical language models: MedGemma (Sellergren et al., 2026), OpenBioLLM, BioMistral (Labrak et al., 2024), and MedPhi. Two larger systems are run alongside them under zero-shot prompting, MedGemma-27B and Gemini 2.5 Flash, the latter a general-purpose model reached through the Google AI API rather than an openly distributed checkpoint. Exact checkpoints for all six are listed in Appendix G.5. Unless otherwise specified, all models are evaluated with identical decoding parameters and without any task-specific fine-tuning. We report accuracy (%) for multiple-choice question answering.

We consider four inference regimes: (1) Zero-Shot (prompt-only), (2) Few-Shot (three in-context examples), (3) Chain-of-Thought (CoT) prompting without external knowledge, and (4) OntoDiagnose, which injects ontology-derived evidence chains into the prompt without modifying model parameters.

4.2 Datasets and Evaluation Protocol

To assess both in-domain performance and cross-dataset generalization, we evaluate OntoDiagnose

on a diverse suite of medical question answering benchmarks: **MedQA (USMLE-style)** (Jin et al., 2020), which consists of professional-level clinical vignettes requiring multi-hop diagnostic reasoning; **MedMCQA** (Pal et al., 2022), a large-scale medical entrance examination dataset emphasizing factual and clinical knowledge; **PubMedQA** (Jin et al., 2019), containing biomedical yes/no/maybe questions grounded in scientific abstracts; **MMLU-Medical** (Hendrycks et al., 2020), comprising medical subsets of MMLU covering anatomy, clinical knowledge, genetics, and professional medicine; and **MedXpertQA** (Zuo et al., 2025), an out-of-distribution benchmark designed to test robustness under distribution shift.

MedQA serves as the primary anchor dataset, while the remaining benchmarks are used to evaluate generalization and robustness. Across all five benchmarks we apply the same prompting, decoding, and option-scoring protocol used for MedQA, so that the with- and without-OntoDiagnose columns are directly comparable.

4.3 Main Results on MedQA

Table 2 reports accuracy on MedQA across models and prompting regimes. The OntoDiagnose column augments zero-shot prompting with ontology-derived evidence chains and leaves the prompt otherwise unchanged, so the gap between it and the zero-shot column isolates the contribution of the evidence itself. OntoDiagnose consistently outperforms prompt-only baselines, yielding an average absolute improvement of +9.64 percentage points over zero-shot inference. The two larger models at the foot of the table were run zero-shot only and are excluded from that average; their size-stratified results appear in Appendix D.1.

4.4 Cross-Dataset Generalization

To evaluate robustness beyond USMLE-style questions, we assess OntoDiagnose on additional medical QA benchmarks. Table 3 reports accuracy across datasets for all models under zero-shot prompting with and without OntoDiagnose, so each pair of rows shows the effect of adding evidence chains while holding the model and prompt fixed.

OntoDiagnose yields consistent improvements across all benchmarks and models. Mean gains are comparable across the four benchmarks, at +6.5 points on MedMCQA, +6.1 on MedXpertQA, +6.1 on PubMedQA, and +4.9 on MMLU-Medical. All four are smaller than the +9.6 points obtained on MedQA, which is consistent with MedQA's vignette-style items being the setting richest in the symptom, finding, and disorder relations that SNOMED CT encodes.

Model	MedMCQA	PubMedQA	MMLU	MedXpertQA
MedGemma	48.37	62.54	55.21	44.73
+ OntoDiagnose	**55.29**	**69.47**	**61.58**	**51.92**
OpenBioLLM	45.18	60.11	51.74	42.18
+ OntoDiagnose	**51.72**	**66.34**	**57.62**	**48.77**
MedPhi	42.23	56.02	51.49	37.91
+ OntoDiagnose	48.07	61.15	53.31	41.86
BioMistral	42.89	58.16	48.67	37.06
+ OntoDiagnose	49.37	64.08	54.12	43.61

Table 3: Generalization across four further medical QA benchmarks (accuracy %). For each model, the first row is zero-shot prompting and the shaded row adds OntoDiagnose. Bold marks the best.

Ontology-grounded evidence therefore appears to reduce reliance on surface-level pattern matching without eliminating the dependence on ontology coverage.

4.5 Comparison with KG-RAG Approaches

We compare OntoDiagnose with representative knowledge-graph-augmented retrieval approaches to contextualize its performance. Table 4 contrasts OntoDiagnose with document-based RAG and KG-RAG-style pipelines using MedQA accuracy under matched conditions (all methods use MedGemma-4B as the backbone). OntoDiagnose outperforms both document-based RAG and KG-RAG, at 61.19% against 52.31% and 54.01%.

We note that this comparison varies both the method and the knowledge source, since our KG-RAG baseline follows prior work in operating over PrimeKG rather than SNOMED CT. It therefore compares two complete pipelines rather than isolating the compression step, and the within-SNOMED controls reported in the same table separate the two effects.

The two control rows inject the same 142-token budget as OntoDiagnose but carry evidence that is either structurally similar and semantically unrelated, or topically medical and independent of the question; neither recovers as much as a single point, which places the gain with question-specific ontology structure rather than with added context. Full details of the KG-RAG setup, including entity extraction, retrieval configuration and serialization format, are given in Appendix F.2, and the full control protocol in Appendix G.6.

Method	Knowledge Source	Evidence Form	Avg. Evidence Tokens	Accuracy (%)
Prompt-only	–	None	0	50.92
Random ontology chains	Any Ontology	Unrelated paths	142	51.38
Generic medical text	Textbook	Unrelated text	142	51.74
Document RAG	Text	Passages	287	52.31
KG-RAG	PrimeKG	Graph paths	287	**54.01**
OntoDiagnose	**SNOMED CT**	**Grounded chains**	**142**	**61.19**

Table 4: Comparison with retrieval-augmented methods and matched-length controls on MedQA. All methods use MedGemma-4B as the backbone. Evidence tokens report the average number of retrieved or serialized evidence tokens per question. The shaded row denotes OntoDiagnose.

4.6 Analysis and Clinical Implications

Few-shot and CoT prompting: Few-shot prompting yields marginal gains, most likely because a handful of demonstrations cannot cover the diversity of clinical presentations. CoT prompting provides modest improvements but does not by itself prevent medically incorrect reasoning in the absence of structured grounding.

Impact of ontology grounding: Gains track baseline accuracy rather than parameter count. Relative to its own zero-shot score, BioMistral improves by 33%, MedGemma-4B by 20%, OpenBioLLM by 20%, and MedPhi by 18%, while MedGemma-27B improves by 10% and Gemini 2.5 Flash by 4%. Structured biomedical knowledge thus appears to substitute for parametric medical knowledge where that knowledge is thinnest, while still adding a smaller amount for stronger models.

5 Discussion

Our experiments demonstrate that an ontology-grounded reasoning scaffold (ONTODIAGNOSE) substantially improves multiple-choice medical question answering across a range of medical language models (see Tables 2, 3, and 4 for core results). These gains are consistent across datasets, model sizes, and prompting regimes. The benefits therefore arise from structured knowledge grounding rather than from prompt engineering alone. All four MedQA improvements are statistically significant at $p < 0.001$ under bootstrap resampling with 10,000 iterations; 95% confidence intervals are reported in Appendix G.3.

Why structured evidence helps: We suggest three contributing factors. First, **grounded retrieval constrains the inference space**: ontology paths explicitly encode multi-hop clinical relations (e.g., symptom → finding → disease) (Lewis et al., 2021; Singhal et al., 2022), which conditions the LM on compact, clinically relevant facts. Second, **spanning-tree compactness reduces noise**: the MST summarization compresses the explored neighborhood into a minimal connector, suppressing generic hub nodes; removing this step degrades accuracy by 3.23 pp (Appendix E). Third, **entity linking anchors the expansion**: our linking stage achieves 84.7% top-1 accuracy (Appendix E), so most seeds are correct, though linking errors remain the largest single failure category at 31% of analyzed errors.

Scale and cost: Gains are largest where baseline accuracy is lowest: BioMistral improves by 33% relative to its own zero-shot score against 20% for MedGemma-4B, even though BioMistral is the larger of the two. Parameter count is therefore the wrong axis: what predicts the gain is how much headroom the model has. Gains diminish but persist for larger models: MedGemma-27B gains +5.98 pp and Gemini 2.5 Flash gains +2.76 pp (Appendix D.1). The ontology pipeline adds only 37–67 ms per query and a fixed 2.5 GB memory overhead (Appendices D.2, D.3), which suggests manageable overhead in the single-A100 setup we measured; suitability for constrained hardware remains untested.

Human evaluation and interpretability: Two board-certified physicians evaluated 100 MedQA cases (Table 5); OntoDiagnose achieved 78.3% fully-supported path rate, 91.5% safe-answer rate, and mean usefulness of 3.7/5 (Cohen's $\kappa = 0.81$). The two margins differ in size: evidence chains are rated fully supported far more often than KG-RAG chains, by 16.2 points, while the gap on answer safety is a narrower 5.3 points (Appendix F.1). Evidence chains are interpretable because each statement maps directly to a traversal path through

named SNOMED CT concepts with explicit relation labels, allowing clinicians to trace the reasoning from observed findings to candidate diagnoses. Appendix H provides detailed qualitative case studies, including an example where the baseline hallucinates an incorrect diagnosis (primary ciliary dyskinesia) while OntoDiagnose correctly converges on cystic fibrosis by surfacing multi-system evidence chains.

Robustness checks: We validate that gains arise from structured ontology grounding rather than confounds: token-length controls (random chains, generic sentences, matched-length document RAG) all fail to approach OntoDiagnose's +10.27 pp gain using the same 142-token budget (Appendix G.6), adding answer options to entity extraction does not improve accuracy (61.03 against 61.19), so option leakage does not drive the reported gains (Appendix G.1), and gains persist at +9.06 pp on the partition where no answer-option surface form appears in the evidence (Appendix H.6). Rewriting the prompt does not move the result either: across five template variants that differ in instruction phrasing, evidence placement and formatting, OntoDiagnose accuracy varies by a standard deviation of 0.63 pp (Appendix G.4). The evidence is also compact rather than merely long: an artificially imposed 50-token allowance, roughly one or two chains, already recovers +5.26 pp, and accuracy saturates near the natural 142-token budget (Appendix H.5). OntoDiagnose is partially additive with CoT (+1.22 pp) and self-consistency decoding (Appendix G.8). Undirected traversal maximizes path coverage (94.1%), while uniform edge weights are retained for transparency (relation-type weighting yields only +0.48 pp; Appendix G.7). Failure analysis identifies entity linking errors (31%), ontology gaps (24%), evidence under-utilization (22%), and hub contamination (14%) as dominant error categories (Appendix F.3).

Scope of our claims: OntoDiagnose builds an ontology-grounded relational scaffold *outside* the language model and supplies it before answer selection. We do not claim to expose or alter the model's internal reasoning: the improvements we report are consistent with better-constrained conditioning, and the controls above are designed to establish that the constraint is question-specific and structural rather than a token-count or lexical-overlap artifact. Our empirical claims are likewise bounded by their setting. They concern multiple-choice medical question answering, openly available medical language models, and a professionally curated clinical ontology. Whether the same pattern transfers to open-ended clinical reasoning, to proprietary frontier systems, or to noisier and automatically constructed knowledge graphs is untested here, and we treat each as open.

Metric	OntoDiagnose	KG-RAG	Prompt-only
Fully supported (%)	**78.3**	62.1	–
Safe answer (%)	**91.5**	86.2	82.7
Diagnostic usefulness (1–5)	**3.7**	2.9	–
Inter-annotator agreement (κ)	**0.81**	0.76	0.72

Table 5: Clinician evaluation of 100 MedQA cases by two board-certified physicians. Agreement is measured using Cohen's κ. Rating scales and annotator commentary are provided in Appendix F.1.

6 Conclusion

ONTODIAGNOSE grounds medical question answering in SNOMED CT without updating model parameters, providing the language model with compact evidence chains that can be checked against the ontology. Accuracy improves across every model and benchmark pairing evaluated, with the largest relative gains for lower-scoring models: BioMistral gains 11.5 percentage points on MedQA, a 33% relative improvement, compared with 4% for the strongest model evaluated. This pattern suggests that structured relational evidence can complement parametric knowledge most where baseline performance is lower.

The evidence remains inspectable after inference. In the clinician audit, 78.3% of OntoDiagnose chains were rated as fully supporting the selected answer, compared with 62.1% for KG-RAG, while 91.5% of OntoDiagnose answers were rated safe compared with 82.7% for prompting alone. Taken together, these results support pairing curated ontologies with language models in the medical QA settings studied here, while leaving limitations in ontology coverage and graph selection unresolved. Extending the framework with complementary knowledge sources and stronger path-selection methods is a natural next step.

7 Limitations

Our results are bound to a professionally curated ontology and to one particular way of walking it. SNOMED CT's editorial quality is a precondition of the method rather than an incidental choice, and

its coverage is still incomplete for rare, emerging, or highly contextual concepts; we have not tested whether the pipeline survives on noisier or automatically constructed graphs, and the KG-RAG comparison in Section 4 varies the method and the knowledge source together, so it isolates neither. Two choices in the traversal are deliberate simplifications. Treating edges as undirected yields an evidence chain for 94.1% of questions against 78.6% under strict direction-preserving traversal, but a minority of chains then cross a non-hierarchical relation against its authored direction, which can read as clinically dubious even when the answer is correct; the clinician audit bounds how often this matters rather than preventing it, since roughly one chain in five was not rated fully supported. Uniform edge weights are likewise a transparent baseline rather than an optimum, though relation-type weighting recovers only +0.48 pp (Appendix G.7). The pipeline carries no patient-specific state either: comorbidities, contraindications, and demographic modifiers, which recent work (Parekh et al., 2026) shows to matter clinically, never enter the traversal. Complementary graphs such as PrimeKG, the hybrid traversal variant where semantic fidelity is critical, and condition-gated edge weighting are the obvious next steps.

Entity linking is the single largest source of downstream error, accounting for 31% of the failures we analyzed and concentrated in abbreviations and polysemous terms. The pipeline keeps top-k candidate lists rather than committing to one concept, and its modular design accommodates stronger linkers such as SNOBERT, cross-encoder reranking, or linker ensembles. We evaluate none of these, because each adds supervision or latency and the numbers we report are those of the training-free configuration. Calibrating grounding confidence and propagating it into path scoring, so that uncertain seeds contribute less to a chain, is the direction we would take first.

Every result here is multiple-choice, scored by log-probability over option tokens. That protocol supports controlled comparison across prompting regimes, but it rewards discrimination among supplied options rather than open-ended clinical reasoning, and it does not test whether a model can produce a correct differential unaided; the benchmarks likewise capture none of the longitudinal, multimodal, or interactive character of real decision-making. Our clinical validation is correspondingly narrow, at 100 cases rated by two physicians, and was sized to ask whether evidence chains are supported, useful, and safe on a focused sample. It does not support subgroup analysis, specialty-stratified conclusions, or any claim about deployment readiness, and multi-institutional validation would be required before one could be made. The gains are also tied to the models and deployment assumptions we tested, and later foundation models may shift the cost-benefit balance of explicit ontology grounding. Free-text differential generation with clinician-rated rubrics, harm auditing, and evaluation across model scales and low-resource settings would address these gaps.

8 Ethical Considerations

OntoDiagnose is a decision-support aid, and final medical judgment must remain with a qualified clinician. Its main ethical advantage over a purely parametric model is that the support for an answer is written down: each chain is a named path through SNOMED CT, so it can be inspected, disputed, or audited. Transparency does not by itself neutralize bias, since ontology coverage and model training data may each serve some populations better than others, and we would expect subgroup analyses and harm-focused evaluation before any deployment. The benchmarks used here consist of publicly released examination and literature questions that carry no patient-identifying information, and the clinician audit was carried out on those same public items; evaluations elsewhere should use de-identified or appropriately licensed data, and system outputs should not expose patient identifiers. One practical constraint follows from the ontology itself: SNOMED CT is distributed under license by SNOMED International, and any index, subgraph, or serialized evidence derived from it inherits those terms, so such material can be exchanged only between parties that already hold a valid license. AI assistants were used for writing support and sentence polishing.

Acknowledgements

We gratefully acknowledge the Koita Centre for Digital Health, the IIT Bombay Trust Lab, and the Department of Computer Science and Engineering at IIT Bombay for the support, resources, and funding that enabled this work. Funding for this project was provided by the IIT Bombay Trust Lab, and we are sincerely grateful for their generous support.

References

Shayantan Banerjee, Pratik Jalan, Rudrapratap Chinhara, Chinmayee Kalle, Pramod P Wangikar, and Kshitij Jadhav. 2026. Ontology-guided harmonization enables unified discovery of public metabolomics studies within and across repositories. *bioRxiv*, pages 2026–07.

Payal Chandak, Kexin Huang, and Marinka Zitnik. 2022. Building a knowledge graph to enable precision medicine. *bioRxiv*.

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

Shaker El-Sappagh, Francesco Franda, Farman Ali, and Kyung-Sup Kwak. 2018. Snomed ct standard ontology based on the ontology for general medical science. *BMC medical informatics and decision making*, 18(1):76.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Fernando Gallego, Guillermo López-García, Luis Gasco-Sánchez, Martin Krallinger, and Francisco J. Veredas. 2024. Clinlinker: Medical entity linking of clinical concept mentions in spanish. *Preprint*, arXiv:2404.06367.

Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh V. Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. 2024. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. *Preprint*, arXiv:2402.07630.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Pratik Jalan, Mukul Joshi, Akhilesh Magotra, and Kshitij Sharad Jadhav. 2026. Medbench: Deliberative evaluation of medical language models. In *BioNLP 2026*, pages 981–991.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Preprint*, arXiv:2009.13081.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2567–2577.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11).

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE transactions on big data*, 7(3):535–547.

Jonathan Kim, Anna Podlasek, Kie Shidara, Feng Liu, Ahmed Alaa, and Danilo Bernardo. 2025. Limitations of large language models in clinical problem-solving arising from inflexible reasoning. *Scientific reports*, 15(1):39426.

Lawrence Kou, George Markowsky, and Leonard Berman. 1981. A fast algorithm for steiner trees. *Acta informatica*, 15(2):141–145.

Mikhail Kulyabin, Gleb Sokolov, Aleksandr Galaida, Andreas Maier, and Tomas Arias-Vergara. 2024. Snobert: A benchmark for clinical notes entity linking in the snomed ct clinical terminology. *Preprint*, arXiv:2405.16115.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *Preprint*, arXiv:2402.10373.

Chaerin Lee, Sohee Park, Hyunsik Na, and Daseon Choi. 2026. Regram: Region-first knowledge graph reasoning for medical question answering. *Preprint*, arXiv:2601.09280.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing. 2024. Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. *Preprint*, arXiv:2305.13269.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns*, 5(3).

Zhenxi Lin, Ziheng Zhang, Xian Wu, and Yefeng Zheng. 2024. Biomedical entity linking as multiple choice question answering. *Preprint*, arXiv:2402.15189.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. *Preprint*, arXiv:2010.11784.

Kurt Mehlhorn. 1988. A faster approximation algorithm for the steiner problem in graphs. *Information Processing Letters*, 27(3):125–128.

Saeel Sandeep Nachane, Ojas Gramopadhye, Prateek Chanda, Ganesh Ramakrishnan, Kshitij Sharad Jadhav, Yatin Nandwani, Dinesh Raghu, and Sachindra Joshi. 2026. Few shot chain-of-thought driven reasoning to prompt llms for open ended medical question answering. *Preprint*, arXiv:2403.04890.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *Preprint*, arXiv:2311.16452.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Jash Rajesh Parekh, Wonbin Kweon, Joey Chan, Rezarta Islamaj, Robert Leaman, Pengcheng Jiang, Chih-Hsuan Wei, Zhizheng Wang, Zhiyong Lu, and Jiawei Han. 2026. Condition-gated reasoning for context-dependent biomedical question answering. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 11774–11785.

Dattaraj J. Rao, Shraddha S. Mane, and Mukta A. Paliwal. 2022. Biomedical multi-hop question answering using knowledge graph embeddings and language models. *Preprint*, arXiv:2211.05351.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 62 others. 2026. Medgemma technical report. *Preprint*, arXiv:2507.05201.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, and 11 others. 2022. Large language models encode clinical knowledge. *Preprint*, arXiv:2212.13138.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, and 12 others. 2023. Towards expert-level medical question answering with large language models. *Preprint*, arXiv:2305.09617.

Karthik Soman, Peter W Rose, John H Morris, Rabia E Akbas, Brett Smith, Braian Peetoom, Catalina Villouta-Reyes, Gabriel Cerono, Yongmei Shi, Angela Rizk-Jackson, and 1 others. 2024. Biomedical knowledge graph-optimized prompt generation for large language models. *Bioinformatics*, 40(9):btae560.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, and 1 others. 2023. Evaluating large language models on medical evidence summarization. *NPJ digital medicine*, 6(1):158.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Zihan Wang, Hao Wang, Shi Feng, Xiaocui Yang, Daling Wang, Yiqun Zhang, Jinghao Lin, Haihua Yang, and Xiaozhong Ji. 2026. Deepmed: Building a medical deepresearch agent via multi-hop med-search data and turn-controlled agentic training & inference. *Preprint*, arXiv:2601.18496.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jiageng WU, Xian Wu, and Jie Yang. 2024. Guiding clinical reasoning with large language models via knowledge seeds. *Preprint*, arXiv:2403.06609.

Wenhao Wu, Zhentao Tang, Yafu Li, Shixiong Kai, Mingxuan Yuan, Zhenhong Sun, Chunlin Chen, and Zhi Wang. 2026. From conflict to consensus: Boosting medical reasoning via multi-round agentic rag. *Preprint*, arXiv:2603.03292.

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

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D Manning, Percy S Liang, and Jure Leskovec. 2022. Deep bidirectional language-knowledge graph pretraining. *Advances in Neural Information Processing Systems*, 35:37309–37323.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 535–546.

Hongyi Yuan, Zheng Yuan, and Sheng Yu. 2022. Generative biomedical entity linking via knowledge base-guided pre-training and synonyms-aware fine-tuning. *Preprint*, arXiv:2204.05164.

Cyril Zakka, Akash Chaurasia, Rohan Shad, Alex R. Dalal, Jennifer L. Kim, Michael Moor, Kevin Alexander, Euan Ashley, Jack Boyd, Kathleen Boyd, Karen Hirsch, Curt Langlotz, Joanna Nelson, and William Hiesinger. 2023. Almanac: Retrieval-augmented language models for clinical medicine. *Preprint*, arXiv:2303.01229.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D. Manning, and Jure Leskovec. 2022. Greaselm: Graph reasoning enhanced language models for question answering. *Preprint*, arXiv:2201.08860.

Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. *Preprint*, arXiv:2305.03268.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *Preprint*, arXiv:2501.18362.

A Pipeline Implementation Details

A.1 Subgraph Summarization Algorithm (MST-on-Closure)

Algorithm 1 gives the summarization procedure referenced in Section 3.4. It takes the ontology graph, the grounded seed set and the traversal radius, computes shortest paths between every pair of seeds, builds a complete closure graph over the seeds weighted by those distances, takes a minimum spanning tree of that closure, and finally expands each tree edge back into the underlying path. The result is a compact tree that connects all seeds while discarding branches that link none of them.

Algorithm 1 Baseline subgraph summarization. The procedure builds a metric closure over the grounded seed nodes, takes a minimum spanning tree, and expands its edges back into ontology paths.

Require: Ontology graph $G = (V, E)$, seed set S, BFS radius r

$V_S \leftarrow N_r(S)$ ▷ Bounded BFS neighborhood around seed nodes

for all $u \in S$ **do** ▷ Compute shortest paths from each seed node

  Run BFS on $G[V_S]$ from u to compute $\mathrm{dist}(u, \cdot)$ and parent pointers

end for

Build closure graph $M = (S, E_M)$ where edge weights $w_{u,v} = \mathrm{dist}(u, v)$

Compute MST: $T_M \leftarrow \mathrm{MST}(M)$

Expand MST edges to real paths: $T \leftarrow \bigcup_{(u,v)\in E(T_M)} p(u, v)$

return compact tree T

A.2 Prompt Template

The following template is used for zero-shot OntoDiagnose inference. Option letters run to the number of choices the benchmark provides, which is four to five across MedQA, MedMCQA and MMLU-Medical. Evidence chains are serialized as numbered statements and placed between the clinical context and the question:

```
Context:
[CLINICAL CONTEXT]
Evidence 1: <E1>
Evidence 2: <E2>
...

Question [Q] with Options [A, B, C, ...]
Task: Select the appropriate answer.
```

For the few-shot setting, three exemplar question-answer pairs are prepended before the clinical context. For CoT + OntoDiagnose, the instruction "Think step by step using the provided evidence" is appended to the task description. Prompt sensitivity analysis across five template variants shows that OntoDiagnose gains are robust (standard deviation 0.63 pp; see Appendix G.4).

B Extended Related Work

Medical LLMs and benchmarks: Singhal et al. (2022) demonstrated that instruction-tuned medical LLMs such as Med-PaLM encode substantial clinical knowledge and perform competitively on aggregated benchmarks, while still lagging behind clini-

cians on safety, robustness, and completeness. Empirical analyses consistently report failure modes including hallucinations, unsupported assertions, and difficulty with multi-hop clinical reasoning under distribution shift (Liévin et al., 2024; Tang et al., 2023; Nachane et al., 2026; Nori et al., 2023). Benchmark datasets including MedQA (USMLE-style) and MultiMedQA have become standard testbeds for evaluating professional-level clinical reasoning (Jin et al., 2020; Singhal et al., 2022).

Medical RAG and structured reasoning: Xiong et al. (2024) systematically benchmarked RAG systems for medicine and found that retriever quality and corpus relevance are primary determinants of RAG effectiveness. Almanac (Zakka et al., 2023) demonstrated that coupling retrieval with curated medical knowledge bases improves factuality over parametric-only inference. Verify-and-Edit (Zhao et al., 2023) introduced a knowledge-enhanced chain-of-thought framework that post-hoc verifies reasoning steps against external evidence, while Chain-of-Knowledge (Li et al., 2024) grounds LLM reasoning in heterogeneous knowledge sources.

Knowledge graphs and ontologies: Biomedical KGs and ontologies such as SNOMED CT (El-Sappagh et al., 2018), Unified Medical Language System (UMLS), and PrimeKG (Chandak et al., 2022) encode curated clinical entities and relations. Prior work in biomedical KG-QA has shown that graph traversal, subgraph extraction, and path-based reasoning can support multi-hop inference (Rao et al., 2022). Ontologies like SNOMED CT play a central role in standardizing clinical concepts and have been widely used for grounding clinical text in decision support systems.

Hybrid pipelines: KG-RAG methods integrate structured knowledge graphs into LLM prompting pipelines by retrieving triples or paths and serializing them as textual context (Soman et al., 2024; Edge et al., 2024). QA-GNN (Yasunaga et al., 2021) jointly reasons over text and knowledge graphs via message passing, while GreaseLM (Zhang et al., 2022) and its successor (Yasunaga et al., 2022) fuse deep LM representations with graph structure. Soman et al. (2024) proposed a KG-RAG pipeline augmenting LLM prompts with biomedical KG triples, and G-Retriever (He et al., 2024) extends this to general graph-structured retrieval. However, most KG-RAG pipelines employ generic traversal without clinically constrained evidence structures.

Entity linking methods: SapBERT (Liu et al., 2021) provides strong embedding-based normalization for biomedical concepts, and recent work such as ClinLinker (Gallego et al., 2024) and generative entity linking methods (Yuan et al., 2022) have explored cross-encoder reranking and retrieval-augmented linking strategies. SNOBERT (Kulyabin et al., 2024) is a two-stage BERT-based system and benchmark for linking clinical-note spans to SNOMED CT; on our evaluation set it reaches +1.6 pp top-1 accuracy over SapBERT (Table 27). Lin et al. (2024) instead cast entity linking as multiple-choice question answering, which gives comparable accuracy at higher latency. While we adopt SapBERT for its favorable accuracy-latency trade-off, integrating specialized EL pipelines could further reduce entity linking errors, the primary failure mode (31% of errors; Appendix F.3).

Concurrent KG reasoning methods: ReGraM (Lee et al., 2026) constructs localized KG "regions" around question entities and constrains inference within region boundaries, sharing the spirit of OntoDiagnose's bounded traversal but using a different region-first construction and no metric-closure compression or serialization step. Parekh et al. (2026) introduce condition-gated edges that model contraindications and patient-specific modifiers, an important clinical dimension that OntoDiagnose does not currently encode (see Section 7). Multi-round agentic retrieval approaches (Wu et al., 2026; Wang et al., 2026) improve robustness through iterative retrieval loops at the cost of increased complexity and latency. OntoDiagnose offers a lightweight, single-pass alternative; future work could integrate it as the structured retrieval component within agentic loops. Med-PaLM 2 (Singhal et al., 2023) demonstrates large gains via instruction fine-tuning and ensemble refinement; OntoDiagnose is orthogonal as a training-free augmentation.

C Entity Grounding and Ontology Exploration Details

Clinical parsing and entity extraction: The first stage of OntoDiagnose handles abbreviations and non-canonical clinical phrasing during clinical text preprocessing and medical named-entity extrac-

tion. Clinical vignettes often contain long free-text descriptions, abbreviations, and non-canonical expressions. To address this, we adopt a hybrid extraction strategy that combines high-precision lexical matching with neural biomedical encoders.

Pipeline:

1. **Segmentation and tokenization**: Sentence segmentation and tokenization are performed using SciSpaCy (Neumann et al., 2019), which is optimized for biomedical and clinical text.
2. **Pre-matching**: Dictionary- and embedding-based matching is applied to capture exact surface forms and common clinical abbreviations with high precision.
3. **Span proposal**: Mentions missed by lexical matching are proposed by a neural biomedical named-entity recognizer running over the SciSpaCy parse (Neumann et al., 2019). SapBERT enters at the linking stage, where it represents the proposed spans; it is not used to detect them.
4. **Normalization**: Candidate mentions are normalized and deduplicated via lowercasing, punctuation removal, and abbreviation expansion.

Lexical and semantic matching provides high-precision detection for canonical clinical terms (e.g., **"optic neuritis", "Lhermitte's sign"**), while the neural component captures context-dependent or non-exact mentions (e.g., **"blurred vision"** $\rightarrow$ **"visual disturbance"**). This hybrid design balances precision and recall while minimizing spurious extractions.

Concept linking / grounding: Each extracted entity $e \in \mathcal{E}$ must be mapped to one or more candidate ontology nodes in V. We follow an embedding + nearest-neighbor linking approach.

Indexing: We precompute SapBERT embeddings $\phi(v) \in \mathbb{R}^d$ for ontology concepts, including canonical names and synonyms, following standard biomedical entity linking practice (Liu et al., 2021). These embeddings are stored in an approximate nearest-neighbor (ANN) index using FAISS to enable scalable similarity search over large concept inventories (Johnson et al., 2019).

Candidate retrieval: Given an extracted entity mention e, we compute its contextual embedding $\phi(e)$ using SapBERT and retrieve the top-k nearest ontology concepts via ANN search:

$$\text{candidate}(e) = \text{ANN-Search}(\phi(e), k).$$

We set $k = 10$ to prioritize high recall during candidate generation.

Disambiguation: Ambiguity among candidate concepts is resolved using a lightweight, rule-guided disambiguation strategy:

1. **Contextual reranking:** candidates are scored by cosine similarity between $\phi(v)$ and a context embedding derived from the surrounding sentence and question stem.
2. **Type constraints:** semantic type compatibility is enforced (e.g., a mention describing a clinical sign is not linked to a pharmacological concept).
3. **LM-based reranking (optional):** the implementation also supports ranking the top-k candidates with a small language model prompt scored on contextual plausibility; this step is applied sparingly to control latency and cost.

Outputs: The linking stage produces a set of ontology-grounded seed nodes

$$S = \bigcup_{e \in \mathcal{E}} \text{top1}(\text{candidate}(e)),$$

which serves as the entry point for subsequent ontology exploration. Top-k candidate lists are optionally retained to support soft grounding and robustness analyses. Embedding-based biomedical entity linking with SapBERT has been shown to provide strong performance for clinical concept normalization (Liu et al., 2021; Yuan et al., 2022), while FAISS enables efficient ANN retrieval at scale (Johnson et al., 2019).

Ontology exploration: Starting from the seed set $S \subseteq V$, the objective of ontology exploration is to identify a compact neighborhood and short, clinically meaningful paths that connect observed entities to diagnostically relevant semantic classes (e.g., disorders or investigations). We perform bounded graph traversal to obtain

$$N_r(S) \triangleq \{v \in V \mid \min_{s \in S} \text{dist}_G(s, v) \leq r\},$$

where the radius r controls the extent of expansion and limits the inclusion of low-relevance nodes.

Short-path enumeration: Within the explored neighborhood, we enumerate shortest paths (by hop count) that (i) originate from a seed node $s \in S$ and (ii) terminate at nodes belonging to target semantic categories (e.g., diseases, disorders, tests). Path length is bounded by $\ell \leq L_{\max}$ (with $L_{\max} = 5$ in our experiments), and only the top-P paths per seed are retained to control computational cost. In practice, bounded traversal combined with degree-based pruning yields tractable exploration for sparse biomedical graphs, while preserving clinically relevant multi-hop relations.

Spanning-tree compression: The explored subgraph $\mathcal{G}_S = (V_S, E_S)$, with $V_S = N_r(S)$, can be large and contain many loosely related nodes. To obtain a compact and interpretable connector structure, we compress $\mathcal{G}_S$ into a tree-shaped subgraph T that links the grounded seed nodes while suppressing irrelevant branches. Our goal is not to recover an optimal reasoning graph, but to construct a minimal, human-interpretable structure that preserves salient multi-hop relations. We adopt a standard metric-closure, MST-like approximation commonly used for practical Steiner tree construction in large graphs (Kou et al., 1981; Mehlhorn, 1988). Specifically, we compute shortest-path distances among the seed nodes S, build a complete closure graph over S using these distances, compute a minimum spanning tree on the closure, and then expand its edges back to paths in $\mathcal{G}_S$, with optional light pruning for relevance. This procedure yields a compact connector that provides bounded approximation guarantees while remaining computationally efficient for large ontologies.

1. For each pair of seed nodes $u, v \in S$, compute the shortest-path distance $\mathrm{dist}(u, v)$ and corresponding path $p(u, v)$ in $\mathcal{G}_S$.
2. Construct a closure graph $M = (S, E_M)$ with edge weights $w_{u,v} = \mathrm{dist}(u, v)$.
3. Compute a minimum spanning tree T_M over M (e.g., using Kruskal or Prim).
4. Expand each edge $(u, v) \in T_M$ into its corresponding shortest path $p(u, v)$ and take their union to form T, optionally pruning low-utility leaves.

In our baseline configuration, all ontology edges are assigned uniform weight ($w_e = 1$), such that distances correspond to hop counts. This choice prioritizes structural simplicity and interpretability over fine-grained semantic weighting. This summarization step is a practical compression mechanism rather than a claim of optimal reasoning, and is intentionally kept stable, and training-free.

D Scalability and Resource Efficiency

D.1 Size-Stratified Performance Analysis

To examine how model scale interacts with ontology grounding, we evaluate OntoDiagnose across three model-size tiers on MedQA (Table 6). We additionally include two larger models, MedGemma-27B and Gemini 2.5 Flash, to assess whether the benefits of ontology grounding diminish for frontier-class systems.

Tier	Model	Parameters	Zero-Shot	Onto-Diagnose	Gain
Small	MedPhi	3B	43.96	51.97	+8.01
	MedGemma	4B	50.92	61.19	+10.27
Medium	BioMistral	7B	35.18	46.69	+11.51
	OpenBioLLM	8B	45.01	53.79	+8.78
Large	MedGemma	27B	62.41	68.39	+5.98
	Gemini 2.5 Flash	–	71.53	74.29	+2.76

Table 6: Size-stratified performance on MedQA (accuracy %) across three parameter tiers, with two larger models added to test whether the benefit of ontology grounding survives at scale. Gains are largest for the lowest-scoring models.

Absolute gains do not decline monotonically with parameter count: small-tier models (3–4B) gain +9.14 pp on average and medium-tier models (7–8B) gain +10.15 pp, so the ordering reverses between those two tiers. The decline appears instead against baseline accuracy. Expressed relative to each model's own zero-shot score, gains fall from 33% (BioMistral, 35.18% zero-shot) through 20%, 20%, and 18% for the remaining small and medium models, to 10% for MedGemma-27B and 4% for Gemini 2.5 Flash. Large-tier models gain +4.37 pp in absolute terms. Even Gemini 2.5 Flash, the strongest model we evaluate (71.53% zero-shot), benefits from ontology grounding (+2.76 pp), which is consistent with explicit relational structure carrying information complementary to parametric knowledge.

D.2 Inference Latency Analysis

We measure end-to-end inference latency for each model with and without OntoDiagnose on a standardized hardware setup (single NVIDIA A100 40GB GPU). Table 7 reports mean latency per question over 500 MedQA samples.

Model	Zero-shot (ms)	Onto-Diagnose (ms)	Difference (ms)	Overhead
MedPhi (3B)	187	224	+37	19.8%
MedGemma (4B)	398	451	+53	13.3%
BioMistral (7B)	342	389	+47	13.7%
OpenBioLLM (8B)	421	488	+67	15.9%

Table 7: Inference latency per question on MedQA. The OntoDiagnose pipeline adds 37–67 ms of overhead (13–20%), comprising entity extraction, SapBERT encoding with FAISS lookup, BFS traversal, MST summarization, and evidence serialization.

The ontology pipeline overhead is dominated by the embedding-based entity linking step (SapBERT encoding and FAISS approximate nearest-neighbor search), which is the single largest contributor. Graph traversal and MST summarization account for most of the remainder, and entity extraction and evidence serialization for the smallest shares. All pipeline stages operate on precomputed indices and cached graph structures, so latency scales linearly with the number of extracted entities rather than with ontology size.

D.3 Memory Footprint

OntoDiagnose introduces a fixed memory overhead independent of the backbone LLM (Table 8):

Component	Memory (GB)
SNOMED CT FAISS index (SapBERT embeddings)	2.1
SNOMED CT graph cache (adjacency + metadata)	0.4
Total pipeline overhead	**2.5**

Table 8: Memory footprint of the OntoDiagnose ontology pipeline. The 2.5 GB overhead is constant regardless of the backbone model and is dominated by the FAISS index containing SapBERT embeddings for all SNOMED CT concept preferred terms and synonyms.

This overhead is modest relative to the memory consumed by the backbone language models themselves (e.g., MedGemma-4B requires approximately 8 GB in half-precision, while BioMistral-7B requires approximately 14 GB). The FAISS index can be further compressed using product quantization if memory constraints are severe.

E Ablation Studies and Design Choices

We conduct ablation experiments to isolate the contribution of each pipeline component. All ablations use MedGemma-4B on MedQA unless otherwise specified.

E.1 Seed set size

We vary the number of extracted seed entities per question to assess the sensitivity of OntoDiagnose to entity extraction completeness (Table 9).

Seeds	0	1–2	3–5	6–7	8–10	10+
Accuracy (%)	50.92	54.31	58.47	60.83	61.19	61.24

Table 9: MedQA accuracy vs. number of seed entities. Performance improves substantially from 0 to 6–7 seeds and then flattens: a moderate number of well-grounded entities is enough to capture the diagnostically relevant ontology structure.

Most of the benefit is present once six or seven entities are grounded. The climb from zero seeds to the 6–7 bucket is 9.91 pp, while every bucket above it adds a further 0.41 pp in total and the 10+ bucket sits within 0.05 pp of the default configuration. Extraction therefore has to be good enough to find a handful of the right concepts rather than exhaustive, which is what keeps the entity stage cheap.

E.2 Traversal radius and path length

We ablate the bounded BFS radius r and maximum path length $L_{\max}$ (Table 10).

Radius r	1	2	3	4	5
Accuracy (%)	55.18	59.42	61.19	60.87	59.94
Path length $L_{\max}$	2	3	4	5	6
Accuracy (%)	56.73	59.88	60.61	61.19	60.52

Table 10: Ablation on BFS radius r (top) and maximum path length $L_{\max}$ (bottom). Performance peaks at $r = 3$ and $L_{\max} = 5$; larger values introduce noisy hub nodes and longer paths dilute evidence quality.

Both parameters have an interior optimum. Accuracy climbs 6.01 pp from radius 1 to radius 3 and then gives back 1.25 pp by radius 5, and the path-length curve peaks at $L_{\max} = 5$ before losing 0.67 pp at 6. Expanding further does not add evidence so much as dilute it: wider neighborhoods admit high-degree hub nodes, and longer paths chain together relations that no longer bear on the question.

E.3 Component ablation

We remove individual pipeline components to measure their contribution (Table 11).

Configuration	Accuracy (%)	Change
Full OntoDiagnose	61.19	–
No summarizer (raw subgraph)	57.96	−3.23
No type constraints in EL	59.41	−1.78
String-match-only EL (no SapBERT)	56.12	−5.07
No evidence (prompt-only baseline)	50.92	−10.27

Table 11: Component ablation on MedQA with MedGemma-4B. Each row removes one component while keeping all others intact. The embedding-based entity linking (SapBERT) contributes the largest individual component gain, followed by the MST summarizer.

Replacing embedding-based linking with plain string matching costs 5.07 pp, more than any other single component, which is consistent with entity linking being the dominant error category in our failure analysis. The summarizer is next at 3.23 pp: handing the model the raw explored subgraph rather than its MST connector gives back roughly a third of what the evidence is worth. Semantic type constraints contribute a smaller 1.78 pp. No single component accounts for the full 10.27 pp, so the gain rests on the composition rather than on any one stage.

E.4 Entity linking quality

We evaluate the entity linking module on a held-out set of 500 annotated clinical mentions mapped to SNOMED CT concepts. Table 12 reports top-1 and top-5 accuracy.

Metric	Top-1	Top-5
Accuracy (%)	84.7	93.2

Table 12: Entity linking accuracy on a held-out evaluation set of 500 clinical mentions. Top-1 accuracy (84.7%) indicates the fraction of mentions where the highest-ranked SNOMED CT concept is correct; top-5 accuracy (93.2%) indicates the fraction where the correct concept appears among the top 5 candidates.

The 8.5 pp gap between top-1 and top-5 says that the correct concept is usually retrieved but not always ranked first. That is why the pipeline retains the full top-k candidate list rather than committing to a single concept, and why confidence-aware grounding would be an extension of the existing stage rather than a redesign of it.

F Clinical Evaluation, Error Analysis, and Case Studies

F.1 Clinician Evaluation Study

Protocol: Two board-certified physicians (one internist, one emergency medicine specialist) independently evaluated 100 randomly sampled MedQA cases processed by OntoDiagnose (MedGemma backbone). For each case, annotators were shown the clinical vignette, the OntoDiagnose evidence chains, the model's selected answer, and the gold-standard answer. Aggregate ratings are reported in Table 13. Each annotator rated the following dimensions:

1. **Path support:** Does the evidence chain clinically support the selected answer? (Fully supported / Partially supported / Unsupported)
2. **Answer safety:** Could the selected answer cause patient harm if acted upon? (Safe / Caution required / Harmful)
3. **Diagnostic usefulness:** How useful is the evidence chain for clinical reasoning? (1–5 Likert scale, 5 = highly useful)

Metric	Onto-Diagnose	KG-RAG	Prompt-only
Fully supported (%)	78.3	62.1	–
At least partially supported (%)	91.8	79.4	–
Safe answer (%)	91.5	86.2	82.7
Usefulness (1–5)	3.7	2.9	–
Cohen's κ	0.81	0.76	0.72

Table 13: Clinician evaluation results on 100 MedQA cases. OntoDiagnose evidence chains receive substantially higher clinical support ratings and answer safety scores compared to KG-RAG and prompt-only baselines. Inter-annotator agreement (Cohen's κ) is strong across all conditions.

Qualitative feedback from annotators indicated that OntoDiagnose evidence chains were "concise and clinically relevant," with the ontology-derived paths often matching the reasoning steps that a clinician would follow during differential diagnosis. The most common criticism was occasional inclusion of overly generic intermediate nodes (e.g., "Disorder of body system"), which were rated as unhelpful but not misleading.

F.2 KG-RAG Experimental Details

For the KG-RAG baseline reported in Table 4, we implement a representative knowledge-graph-augmented retrieval pipeline following Soman et al. (2024). Configuration details:

- **Entity extraction:** BERT-based biomedical NER (PubMedBERT fine-tuned on BC5CDR + NCBI-Disease).
- **Knowledge graph:** PrimeKG (Chandak et al., 2022) (129,375 nodes, 4,050,249 relations, counted as 8,100,498 directed edges, across 20

Category	Zero-Shot (%)	Onto-Diagnose (%)	Gain
Pathophysiology / Diagnosis	48.31	60.72	+12.41
Pharmacology / Treatment	52.17	59.28	+7.11
Anatomy / Histology	54.62	62.18	+7.56
Microbiology / Immunology	49.83	58.94	+9.11
Biostatistics / Ethics	55.41	58.63	+3.22
Behavioral Science	51.27	55.89	+4.62

Table 14: Performance by MedQA question category with MedGemma-4B. Gains are largest where multi-hop reasoning over disease-finding relations matters most.

relation types).

- **Retrieval:** Dense retrieval using SapBERT embeddings; top-k=10 triples per entity.
- **Graph traversal:** 3-hop neighborhood expansion from retrieved entities.
- **Serialization:** Retrieved triples serialized as "subject → relation → object" statements, concatenated into the prompt context.
- **Average evidence length:** 287 tokens per question (compared to 142 tokens for OntoDiagnose).

Key differences from OntoDiagnose: (1) KG-RAG uses a general-purpose knowledge graph (PrimeKG) rather than a curated clinical ontology (SNOMED CT); (2) KG-RAG retrieves flat triples without structural summarization, resulting in longer and potentially noisier evidence; (3) OntoDiagnose applies MST-based compression to produce compact, tree-structured evidence chains. For the document-based RAG baseline, we use MedCPT (Jin et al., 2023) to retrieve the top-5 passages from a PubMed abstract corpus, which are concatenated into the prompt context.

F.3 Stratified Error Analysis

We analyze OntoDiagnose performance by MedQA question category to identify areas of strength and weakness (Table 14).

The gain tracks how much of a question SNOMED CT can represent. Pathophysiology and diagnosis items gain most at +12.41 pp, followed by microbiology and immunology at +9.11 pp; these are the categories built from the disorder, finding, and causative-agent relations the ontology encodes densely. Biostatistics and ethics gain least at +3.22 pp and behavioral science at +4.62 pp, since neither turns on the relations SNOMED CT curates. The 9.19 pp spread between the best and worst category is a coverage effect rather than a reasoning one.

Configuration	Accuracy (%)
OntoDiagnose (vignette + stem only)	61.19
OntoDiagnose (vignette + stem + options)	61.03
OntoDiagnose (options only)	53.47
Prompt-only (no evidence)	50.92

Table 15: Answer-option leakage control on MedQA (MedGemma-4B). The four configurations differ only in which text is visible to entity extraction, isolating whether information from the answer options can reach the evidence chains.

Failure mode analysis. Among the 38.81% of questions answered incorrectly by OntoDiagnose + MedGemma, we manually categorized 100 random errors:

- **Entity linking errors** (31%): Incorrect concept canonicalization, particularly for abbreviations (e.g., "MS" linked to "musculoskeletal" instead of "multiple sclerosis") and polysemous terms.
- **Ontology coverage gaps** (24%): Concepts not present in SNOMED CT (e.g., specific drug brand names, emerging biomarkers), or concepts with insufficient relational context.
- **Evidence ignored by LLM** (22%): Correct evidence chains provided but the model selected an incorrect answer, often defaulting to a common-but-wrong pattern from its training distribution.
- **Hub node contamination** (14%): High-degree generic nodes introduced noise despite degree-based pruning, particularly for vignettes involving common symptoms with many differential diagnoses.
- **Other** (9%): Ambiguous questions, annotation disagreements, or edge cases where multiple answers were clinically defensible.

G Robustness and Control Experiments

G.1 Answer-Option Leakage Control

A potential concern with ontology-augmented pipelines is that information from answer options could leak into the evidence generation process, providing the model with an unfair advantage. To address this, OntoDiagnose extracts entities exclusively from the clinical vignette and question stem, explicitly excluding answer-choice text from entity extraction, grounding, and ontology exploration. We validate this design with a controlled experiment (Table 15) comparing four configurations on MedQA (MedGemma-4B):

Letting the extractor see the answer options moves

Relation Type	Authored Direction	Frequency (%)
Is-a (subsumption)	Source → Target	42.3
Finding-site	Source → Target	18.7
Associated-morphology	Source → Target	12.1
Causative-agent	Source → Target	8.9
Has-definitional-manifestation	Source → Target	7.4
Due-to	Source → Target	5.8
Occurrence	Source → Target	2.1
Other (method, laterality, etc.)	Various	2.7

Table 16: SNOMED CT relation types traversed during ontology exploration. The Direction column reports how each relation is authored in SNOMED CT, not the traversal policy used here.

accuracy by −0.16 pp, so nothing useful leaks from the options into the evidence chains. The options-only row is the informative one: with the vignette hidden, evidence built from answer text alone reaches 53.47%, which is 2.55 pp above prompt-only but 7.72 pp below the full configuration. What the evidence contributes is therefore tied to the clinical narrative and not to the candidate answers.

G.2 SNOMED CT Relation Types and Traversal Details

Table 16 lists the SNOMED CT relation types traversed during ontology exploration, along with their authored directionality and their frequency in the explored subgraphs.

High-degree nodes: Nodes whose degree exceeds a threshold δ, set to 500 in our experiments, are treated as hubs and down-weighted by the factor $\min(1, \delta/\deg(v))$ when candidate paths are ranked. A path through a very high-degree node is therefore retained only when no more specific alternative connects the same pair of seeds. Additionally, nodes belonging to the following overly generic SNOMED CT semantic types are pruned prior to summarization: "Clinical finding" (root), "Body structure" (root), "Event" (root), and "Observable entity" (root). This approach reduces hub contamination while preserving clinically specific intermediate nodes.

G.3 Statistical Significance

We assess the statistical significance of OntoDiagnose improvements using bootstrap resampling (10,000 iterations) on MedQA. Table 17 reports 95% confidence intervals for the accuracy difference between OntoDiagnose and the zero-shot baseline.

Model	Accuracy Gain	95% Confidence Interval	p-value
MedGemma (4B)	+10.27	[8.41, 12.13]	< 0.001
OpenBioLLM (8B)	+8.78	[6.92, 10.64]	< 0.001
MedPhi (3B)	+8.01	[6.14, 9.88]	< 0.001
BioMistral (7B)	+11.51	[9.28, 13.74]	< 0.001

Table 17: Bootstrap 95% confidence intervals and p-values for OntoDiagnose improvements over zero-shot baselines on MedQA (10,000 bootstrap iterations). All improvements are statistically significant at $p < 0.001$.

No confidence interval crosses zero, and the narrowest lower bound is +6.14 pp. The intervals for MedGemma-4B, OpenBioLLM, and MedPhi overlap one another, so the paper does not claim a reliable ordering among those three; BioMistral's interval is both the highest and the widest, which is what one would expect of the model with the most headroom and the noisiest baseline.

G.4 Prompt Sensitivity Analysis

To assess robustness to prompt phrasing, we evaluate OntoDiagnose with five prompt template variants on MedQA (MedGemma-4B). Variants differ in instruction phrasing, evidence placement, and formatting, results are shown in Table 18:

Prompt variant	Zero-shot (%)	OntoDiagnose (%)
Default template	50.92	61.19
Evidence before question	50.92	60.87
Bullet-point evidence format	50.92	60.53
Minimal instruction (no task description)	49.18	59.71
Verbose instruction (detailed task description)	51.34	61.42
Mean ± Std.	**50.66 ± 0.81**	**60.74 ± 0.63**

Table 18: Prompt sensitivity on MedQA (MedGemma-4B) across five template variants differing in instruction phrasing, evidence placement and formatting. The final row reports the mean and standard deviation across variants.

The spread across templates is 1.71 pp, from 59.71% under the minimal instruction to 61.42% under the verbose one, and the OntoDiagnose column varies slightly less than the zero-shot column it builds on. Placement and formatting matter least: moving the chains ahead of the question costs 0.32 pp and rendering them as bullets costs 0.66 pp. The gain over zero-shot survives every variant, so it is not an artifact of the particular wording we settled on.

G.5 Reproducibility and Licensing

SNOMED CT licensing: SNOMED CT is maintained by SNOMED International and requires a license for use. Member countries receive the license through their National Release Centers; affiliates of non-member countries may obtain a license through SNOMED International's Affiliate Licensing program. Our experiments use the SNOMED CT International Edition (January 2024 release) (El-Sappagh et al., 2018). Researchers wishing to reproduce this work should obtain an appropriate SNOMED CT license through their institutional or national agreements. Because the ontology itself is licensed, any derived index, subgraph or serialized evidence built from it inherits those licensing terms, and such material can only be shared between parties that already hold a SNOMED CT license.

Software and data: Table 19 lists the exact checkpoint used for each backbone model.

Model	Checkpoint
MedGemma (4B)	`google/medgemma-4b-it`
MedGemma (27B)	`google/medgemma-27b-text-it`
BioMistral (7B)	`BioMistral/BioMistral-7B`
OpenBioLLM (8B)	`aaditya/Llama3-OpenBioLLM-8B`
MedPhi (3B)	`bluesky333/medphi2`
Gemini 2.5 Flash	`gemini-2.5-flash`

Table 19: Model identifiers for the six backbone models evaluated in this work. The five open models are the publicly distributed checkpoints under these identifiers; Gemini 2.5 Flash was accessed through the Google AI API.

All experiments use publicly available model weights accessed through HuggingFace (MedGemma, BioMistral, OpenBioLLM, MedPhi) and Google AI (Gemini). Medical QA benchmarks (MedQA, MedMCQA, PubMedQA, MMLU-Medical) are publicly available. MedXpertQA is accessed through its official release. SciSpaCy, SapBERT, and FAISS are open-source software packages.

Reproducibility details: All models are evaluated with greedy decoding (temperature = 0, top-p = 1.0) to ensure deterministic outputs. For the few-shot setting, we use three fixed exemplars sampled from the MedQA training set. The ontology pipeline uses the following hyperparameters: BFS radius $r = 3$, maximum path length $L_{\max} = 5$, top-k entity linking candidates $k = 10$, and hub node degree threshold $\delta = 500$.

Condition	Accuracy (%)	Evidence Tokens
Prompt-only (no evidence)	50.92	0
Random ontology chains (matched length)	51.38	142
Generic medical sentences (matched length)	51.74	142
Token-matched document RAG	52.19	142
Unmatched document RAG (natural length)	52.31	287
KG-RAG	54.01	287
OntoDiagnose	**61.19**	**142**

Table 20: Control experiments on MedQA with MedGemma-4B. Each control matches the OntoDiagnose token budget but varies whether the evidence is question-specific and structured.

G.6 Control Experiments: Isolating Ontology Grounding Benefits

A key concern is whether OntoDiagnose gains arise specifically from structured ontology grounding or simply from injecting additional tokens into the prompt. We conduct three control experiments on MedQA with MedGemma-4B to disentangle these effects (Table 20): (1) **Random ontology chains**: OntoDiagnose evidence chains are replaced with randomly sampled SNOMED CT paths of equal average length, structurally identical but semantically unrelated to the clinical vignette. (2) **Generic medical sentences**: evidence chains are replaced with matched-length sentences drawn from a medical textbook corpus (Harrison's Principles), selected without regard to question content. (3) **Token-matched document RAG**: document-based RAG passages strictly matched for token count to OntoDiagnose output (142 tokens per question on average).

These results establish that: (1) simply adding tokens to the prompt does not explain OntoDiagnose gains; (2) generic medical knowledge provides negligible benefit without question-specific grounding; and (3) OntoDiagnose achieves roughly eight times larger gains than token-matched document RAG, so structured ontology-derived evidence is the primary driver of improvement.

G.7 Directionality and Edge Weighting Ablation

We ablate edge directionality and weighting choices on MedQA with MedGemma-4B.

Directionality ablation: We compare three traversal regimes: (1) **Undirected** (default): all edges traversed bidirectionally; (2) **Directed-only**:

edges traversed only in their defined direction (e.g., "finding-site" from finding to body structure, "is-a" only upward); (3) **Hybrid**: is-a edges bidirectional, all other edges directed-only. Table 21 reports the comparison.

Traversal	Accuracy (%)	Mean Paths	Mean Hops	Coverage
Undirected (default)	61.19	5.8	3.2	94.1%
Directed-only	58.73	3.1	2.7	78.6%
Hybrid (is-a bidir.)	60.84	4.9	3.0	91.2%

Table 21: Directionality ablation on MedQA, comparing undirected, direction-preserving and hybrid traversal. Coverage is the share of questions receiving at least one evidence chain, and average paths and hops describe the chains produced.

Directed-only traversal sharply reduces path coverage because many clinical reasoning chains require "upward" generalization through the is-a hierarchy followed by "downward" specialization. For example, linking "optic neuritis" to "multiple sclerosis" requires ascending from a specific finding to a broader category and then descending to a disease concept. The undirected default balances coverage and accuracy; the hybrid variant offers a principled middle ground that we recommend for settings where edge semantic fidelity is critical.

Edge weighting ablation: We compare three weighting schemes: (1) **Uniform** (default): all edges weight 1; (2) **Relation-type weighting**: edges weighted by empirically estimated clinical relevance (due-to: 0.5, causative-agent: 0.6, finding-site: 0.8, is-a: 1.0, other: 1.2); (3) **Inverse-frequency weighting**: edge weight = $\log(1 + \text{freq}(r))$ where $\text{freq}(r)$ is the relation type frequency in explored subgraphs. Table 22 reports the comparison.

Weighting	Accuracy (%)	Change
Uniform (default)	61.19	–
Relation-type weighting	61.67	+0.48
Inverse-frequency weighting	61.41	+0.22

Table 22: Edge weighting ablation on MedQA, comparing uniform weights against relation-type and inverse-frequency schemes. Uniform weights are retained as the default for simplicity and reproducibility.

Neither scheme changes the picture. Relation-type weights add 0.48 pp and inverse-frequency weights 0.22 pp, both an order of magnitude below the 10.27 pp the evidence itself contributes, and the relation-type scheme needs clinical relevance values that we set by hand. We therefore keep uniform weights, which leaves hop count as the only quantity the summarizer optimizes and the compression fully determined by the graph rather than by a tuned parameter.

G.8 CoT + OntoDiagnose Combination

We evaluate whether OntoDiagnose benefits are additive with chain-of-thought (CoT) prompting and self-consistency decoding (Wang et al., 2022) on MedQA (Table 23).

Setting	Med-Gemma	Bio-Mistral	Open-BioLLM
Zero-shot	50.92	35.18	45.01
Chain-of-thought	52.12	41.18	46.95
Zero-shot + OntoDiagnose	61.19	46.69	53.79
Chain-of-thought + OntoDiagnose	62.41	48.37	55.12
+ self-consistency (n=5)	63.18	49.21	55.87

Table 23: Combining OntoDiagnose with chain-of-thought prompting and self-consistency decoding on MedQA. Each row adds one component to the row above it.

Chain-of-thought on top of OntoDiagnose adds a further +1.22 pp for MedGemma, +1.68 pp for BioMistral and +1.33 pp for OpenBioLLM, so the two are partially additive rather than redundant, and self-consistency with n=5 adds +0.79 pp on average beyond that. The combination produces the best overall performance by pairing explicit ontology-derived evidence with step-by-step reasoning. However, the CoT contribution is modest compared to the ontology grounding contribution (average +1.41 pp for CoT on top of OntoDiagnose against +9.64 pp for OntoDiagnose alone). Structured evidence is the dominant factor. Self-consistency decoding (Wang et al., 2022) provides additional marginal gains by aggregating over multiple reasoning paths.

H Qualitative Case Studies and Evidence Analysis

We present three worked examples, covering a correct prediction, an incorrect one, and a case where ontology evidence removes a hallucination, followed by analyses of chain count, token budget, and surface-form overlap.

H.1 Case 1: Correct prediction (Cystic Fibrosis)

Question: A 4-year-old boy presents with frequent respiratory tract infections and chronic di-

arrhea. Stools are bulky and greasy with approximately 8 bowel movements daily. He is at the 10th percentile for height and 25th percentile for weight. Chest examination shows intercostal retractions with diffuse wheezing and expiratory rhonchi. What is the most likely cause?
Options: (A) Increased trinucleotide repeats, (B) Defective ciliary protein function, (C) Intracellular retention of misfolded proteins, (D) Altered protease inhibitor configuration, (E) Frameshift mutation of muscle-anchoring proteins.
Gold answer: C

Extracted entities and grounding:

- "frequent respiratory infections" → Recurrent respiratory infection (195839008)
- "chronic diarrhea" → Chronic diarrhea (62315008)
- "bulky greasy stools" → Steatorrhea (271807003)
- "intercostal retractions" → Intercostal recession (609234002)
- "wheezing" → Wheezing (56018004)

Serialized evidence chains (after MST compression):

1. Steatorrhea **is due to** Pancreatic exocrine insufficiency, which **is due to** Cystic fibrosis. Cystic fibrosis causes Mucus gland hyperplasia in the exocrine pancreas.
2. Recurrent respiratory infection **is due to** Cystic fibrosis, which produces Viscous mucus in the Bronchus structure, leading to Bronchiectasis.
3. Wheezing **is due to** Obstruction of airway, which **is due to** Cystic fibrosis and its viscous mucus production.

Model output with OntoDiagnose: "The ontology evidence converges on cystic fibrosis as the unifying diagnosis explaining both GI (steatorrhea via pancreatic insufficiency) and pulmonary (recurrent infections, wheezing via airway obstruction) findings. The molecular mechanism of cystic fibrosis is the CFTR gene mutation causing intracellular retention of misfolded CFTR protein. **Answer: C**."

Model output without OntoDiagnose: MedGemma selected option B (defective ciliary protein function), reasoning about primary ciliary dyskinesia, which explains respiratory infections but not the GI symptoms.

H.2 Case 2: Incorrect prediction (Drug-Induced Lupus)

Question: A 55-year-old man treated with hydralazine for hypertension develops joint pain, pleuritic chest pain, and a malar rash. ANA is positive with antihistone antibodies. What is the most appropriate next step?
Options: (A) Start prednisone, (B) Discontinue hydralazine, (C) Start hydroxychloroquine, (D) Obtain renal biopsy.
Gold answer: B

Serialized evidence chains:

1. Hydralazine **causative-agent of** Drug-induced lupus erythematosus, which **has manifestation** Arthralgia.
2. Drug-induced lupus **has manifestation** Pleurisy AND Malar rash.
3. Drug-induced lupus **is-a** Lupus erythematosus, which **has manifestation** ANA positivity.

Model output with OntoDiagnose: "The evidence chains correctly identify drug-induced lupus caused by hydralazine. The treatment approach for lupus includes immunosuppression. **Answer: A** (Start prednisone)."

Failure analysis: The evidence chains correctly identified the diagnosis (drug-induced lupus) but did not encode the management principle that drug-induced lupus resolves upon discontinuation of the offending agent. This reflects an **ontology coverage gap**: SNOMED CT encodes disease-finding-cause relations but does not explicitly represent first-line management strategies. The model defaulted to its parametric association between lupus and immunosuppressive therapy rather than the more specific management of drug-induced variants.

H.3 Case 3: Hallucination reduction example (Rheumatic Fever)

Question: A 10-year-old boy presents two weeks after a sore throat with migratory polyarthritis, a new cardiac murmur, and subcutaneous nodules. What is the most likely diagnosis?
Options: (A) Juvenile idiopathic arthritis, (B) Acute rheumatic fever, (C) Infective endocarditis, (D) Systemic lupus erythematosus.
Gold answer: B

Model output without OntoDiagnose: "The presentation of migratory polyarthritis and subcutaneous nodules in a child is most consistent with juvenile idiopathic arthritis (JIA). JIA commonly presents with joint involvement and can feature nodules. **Answer: A**."

Hallucination analysis: The baseline model **hallucinated** an incorrect clinical association: while JIA can cause arthritis, the combination of post-streptococcal onset, migratory polyarthritis, cardiac

involvement, and subcutaneous nodules is pathognomonic for acute rheumatic fever (Jones criteria). The model generated a plausible-sounding but factually incorrect reasoning chain, a hallucination pattern where the model relies on surface-level associations (arthritis + child → JIA) rather than structured multi-hop reasoning.

OntoDiagnose evidence chains:

1. "Sore throat" **is due to** Streptococcal pharyngitis; Acute rheumatic fever **is due to** Streptococcal pharyngitis.
2. Acute rheumatic fever **has manifestation** Migratory polyarthritis AND Carditis AND Subcutaneous nodules.

Model output with OntoDiagnose: "The ontology evidence links the prior sore throat to streptococcal infection, which is the causative agent of acute rheumatic fever. The Jones criteria manifestations (migratory polyarthritis, carditis evidenced by the new murmur, subcutaneous nodules) all converge on ARF. **Answer: B**."

This example illustrates how OntoDiagnose reduces hallucinations by providing structured evidence that anchors the model in curated clinical relations rather than allowing unconstrained parametric recall.

H.4 Evidence chain count and length ablation

We ablate the number and average length (in hops) of injected evidence chains per question (Table 24).

Maximum chains	Accuracy (%)	Mean tokens	Evidence ignored
1	56.83	38	8.2%
3	59.94	87	5.1%
5 (default)	61.19	142	3.7%
8	61.07	218	6.4%
12	60.42	311	9.8%

Table 24: Ablation on the number of evidence chains injected per question. The ignore rate is the share of cases whose reasoning text references no provided evidence statement.

Accuracy peaks at the default of five chains, and the ignore rate is lowest there as well, at 3.7%. Both directions away from that point cost something: a single chain loses 4.36 pp because the evidence is often incomplete, while twelve chains lose 0.77 pp and push the ignore rate to 9.8%, so a long evidence block is more likely to be passed over than a short one. The ignore rate is measured on a separate generation run, since the accuracy figures reported throughout the paper come from option scoring rather than free-form output, and a case counts as ignoring the evidence when its reasoning text refers to none of the statements provided.

H.5 Token Budget and Evidence Statistics

Evidence statistics: Across the MedQA evaluation set, OntoDiagnose injects an average of 4.7 evidence chains per question (median: 5, range: 1–5). The average token budget consumed by evidence chains is 142 tokens (median: 134, range: 28–247), which is substantially more compact than KG-RAG and document RAG, which both average 287 tokens per question.

Context limits: We evaluate OntoDiagnose under artificially constrained token budgets (Table 25). Even a severe 50-token allowance, roughly one or two chains, still yields +5.26 pp over prompt-only, and accuracy saturates near the natural 142-token budget with no further benefit beyond 200 tokens. The compression therefore concentrates most of the useful signal into a small number of chains:

Maximum evidence tokens	Accuracy (%)	Mean chains
50	56.18	1.4
100	59.63	3.1
142 (natural, default)	61.19	4.7
200	61.31	5.9
300	61.08	7.2
Prompt-only (0 tokens)	50.92	0

Table 25: OntoDiagnose accuracy under artificially constrained evidence token budgets, from a severe 50-token limit to twice the natural budget. Average chains gives the number of evidence statements that fit within each budget.

H.6 Surface-Form Overlap Analysis

A potential confound is that evidence chains may contain the surface form of the correct answer, trivially revealing the answer through lexical overlap rather than reasoning. We measure and control for this effect.

Measurement: We compute the fraction of MedQA questions where any evidence chain contains a substring match (after lowercasing and stemming) with any of the four to five answer option texts. We then partition questions into "overlap" (evidence contains at least one answer option surface form) and "no overlap" groups, reporting accuracy for each partition in Table 26. Gains persist at +9.06 pp on the no-overlap partition, which covers 58.3% of questions, so lexical overlap is not what drives the improvement. The larger +13.55

pp gain on the partition that overlaps the correct answer is expected, since those questions tend to involve diseases and findings that SNOMED CT covers well.

Partition	Share of questions (%)	Zero-shot (%)	Onto-Diagnose (%)
All questions	100%	50.92	61.19
Overlap with correct answer	23.4%	51.87	65.42
Overlap with any answer	41.7%	51.23	63.18
No overlap with any answer	58.3%	50.71	59.77

Table 26: Surface-form overlap analysis. Questions are partitioned by whether any evidence chain contains the surface form of an answer option, so that gains can be measured separately on the partition where no such lexical cue exists.

Paraphrasing control: To further rule out surface-form confounds, we conduct a control where all evidence chains are paraphrased to replace any verbatim answer-option text with synonymous medical terminology (e.g., "multiple sclerosis" $\rightarrow$ "demyelinating autoimmune disease of the CNS"). Under this condition, OntoDiagnose achieves 60.47% accuracy (-0.72 pp against the default): nearly all of the gain survives when surface-form overlap is eliminated.

I Entity Linking Evaluation Protocol

I.1 Annotation protocol

The held-out entity linking evaluation set (Table 12) consists of 500 clinical mention spans extracted from 100 MedQA vignettes. Two annotators (one medical informaticist, one NLP researcher with biomedical domain expertise) independently mapped each mention to the most appropriate SNOMED CT concept using the SNOMED CT Browser. Mentions were presented in their original vignette context to enable contextual disambiguation.

Annotation guidelines:

1. Each mention is mapped to exactly one SNOMED CT concept (the most specific applicable concept).
2. Mentions that correspond to no SNOMED CT concept are marked as "unmappable" (3.2% of mentions).
3. Mentions that are ambiguous are resolved using the clinical context of the vignette.

Inter-annotator agreement: Cohen's κ for exact concept match was 0.83 (strong agreement). Disagreements (8.4% of mentions) were resolved by a third adjudicator (a physician). Most disagreements involved granularity (e.g., "chest pain" mapped to "Chest pain" vs. "Pleuritic chest pain") rather than semantic category errors.

I.2 Comparison with alternative EL pipelines

We benchmark four entity linking approaches on our held-out evaluation set. Table 27 reports accuracy and per-mention latency for each.

Entity Linking Pipeline	Top-1 (%)	Top-5 (%)	Latency (ms)
String matching only	61.2	71.8	3
SapBERT (ours)	84.7	93.2	14
SNOBERT (Kulyabin et al., 2024)	86.3	94.1	18
Lin et al. (2024)	85.8	93.9	42

Table 27: Entity linking comparison on 500 annotated clinical mentions, reporting top-1 and top-5 accuracy alongside per-mention latency for four linking pipelines, from plain string matching to SNOMED-specific and generative approaches.

SNOBERT's marginal accuracy advantage (+1.6 pp top-1) comes from its SNOMED CT-specific pretraining, which better handles polysemous clinical terms. However, propagating this improvement through the full OntoDiagnose pipeline yields only +0.4 pp end-to-end MedQA accuracy (61.19 $\rightarrow$ 61.58). Downstream MST summarization and answer selection therefore absorb part of the linking error. We retain SapBERT as the default for its lower latency and wider availability.