
Simulating *Viva Voce* Examinations to Evaluate Clinical Reasoning in Large Language Models

Christopher Chiu
Georgia Institute of Technology
cchiu65@gatech.edu

Silviu Pitis
University of Toronto
spitis@cs.toronto.edu

Mihaela van der Schaar
University of Cambridge
mv472@cam.ac.uk

Abstract

Clinical reasoning in medicine is a hypothesis-driven process where physicians refine diagnoses from limited information through targeted history, physical examination, and diagnostic investigations. In contrast, current medical benchmarks for large language models (LLMs) primarily assess knowledge recall through single-turn questions, where complete clinical information is provided upfront. To address this gap, we introduce VivaBench, a multi-turn benchmark that evaluates sequential clinical reasoning in LLM agents. Our dataset comprises 1152 physician-curated clinical vignettes structured as interactive scenarios that simulate a viva voce examination in medical training, requiring agents to actively probe for relevant findings, select appropriate investigations, and synthesize information across multiple steps to reach a diagnosis. We evaluated several state-of-the-art LLMs and found that while models demonstrate competence in diagnosing conditions within well-described clinical presentations, their performance degrades significantly when required to navigate diagnostic uncertainty. Our analysis identified several failure modes that mirror common issues in clinical practice, including: (1) fixation on initial hypotheses, (2) excessive investigation ordering, (3) premature diagnostic closure, and (4) missing critical conditions. These patterns reveal fundamental limitations in how current LLMs manage uncertainty and gather information sequentially. Through VivaBench, we provide a standardized benchmark for evaluating conversational medical AI systems for real-world clinical decision support. Beyond medical applications, we contribute to the larger corpus of research on agentic AI by demonstrating how sequential reasoning trajectories can diverge in complex decision-making environments.

1 Introduction

Many future applications of Large Language Models (LLMs), such as LLM-assisted clinical reasoning and diagnostics, will involve multi-turn, sequential interactions, where an LLM or LLM-based agent will need to actively gather and act on information that only becomes available at inference time. Most current LLM evaluations, however, including those in healthcare [25] and those aspiring to be comprehensive assessments of general intelligence such as “humanity’s last exam” [37, 11], evaluate LLMs based on single-turn interactions [31, 45]. While this approach is valuable for gauging foundational knowledge and specific skills like question answering [24] or knowledge encoding [43], it falls short of capturing the dynamic, iterative nature of complex real-world problem-solving that is so central in critical domains like healthcare.

A growing body of recent work examines the multi-turn conversational and reasoning capabilities of LLMs [21, 49, 50], and observes that even frontier models exhibit significantly lower performance in multi-turn settings and suffer from issues of derailment and error propagation when reasoning over multiple steps [28, 29, 44]. One such setting, and the subject of this work, is that of clinical

decision-making, wherein the derailment of naturally emerging multi-step interactions can have significant consequences for patients.

In clinical practice, the diagnostic journey of a patient can be conceptualized as a Partially Observable Markov Decision Process [46]. Physicians begin with limited, often ambiguous, information (generic findings) and iteratively gather more data through history taking, physical examinations, and diagnostic tests, as is roughly illustrated in Figure 1. Each new piece of information allows them to update their belief state about the patient’s underlying condition, refining their differential diagnosis in a manner akin to Bayesian inference, where the probability of a specific diagnosis given the observed findings, $P(\text{specific diagnosis}|\text{generic findings})$, is continuously updated. This iterative, probabilistic reasoning is fundamental to how clinicians navigate diagnostic uncertainty.

Schubert et al. [41] have explored the concept of a "diagnostic trajectory" in depth, emphasizing the sequential nature of clinical encounters and the accumulation of evidence over time. Their framework highlights the importance of timely and accurate diagnosis, which relies on the ability to effectively navigate these trajectories. This underscores the need for AI models that are not only knowledgeable but are also capable of robust diagnostic reasoning along these evolving pathways. A critical question then arises: how can we effectively measure and benchmark this sophisticated, sequential reasoning capability in AI systems to ensure they are safe and effective for clinical use?

Viva voce examinations A viva voce examination is an interactive oral exam used to test the clinical skills of medical students. While it is most commonly used as thesis defense within academia, it has also been used and validated as an examination tool in medical training [7, 12]. A typical example, as shown in Figure 1, begins with a simulated patient scenario, and the student physician then queries the patient to answer a series of diagnosis / case management questions. The verbal nature of this examination translates nicely, allowing for easy administration to a “student” LLM.

Inspired by this, we propose VivaBench, which operationalizes the viva voce concept for evaluating sequential clinical reasoning capabilities of LLM-based agents. It simulates a multi-turn diagnostic encounter, akin to a medical viva voce examination, where an agent must actively solicit information through history taking, physical examination requests, and ordering of diagnostic investigations to arrive at a diagnosis for a presented clinical case. Our key contributions are:

- VivaBench is an open-source, open-domain dataset with physician-curated structured clinical vignettes. Each vignette is manually reviewed and structured for interactive probing, requiring agents to navigate diagnostic uncertainty from limited initial information. This interactive format moves beyond static Q and A to assess dynamic information gathering and hypothesis refinement.
- We provide an extensible evaluation framework for the interactive administration of VivaBench vignettes, which is designed to provide accurate and deterministic outputs for reproducibility.
- We evaluate 6 leading LLMs and systematically identified critical failure modes. These patterns mirror common cognitive errors in clinical practice and reveal fundamental limitations in how current LLMs manage uncertainty and synthesize information sequentially in high-stakes scenarios, including: (1) fixation on initial hypotheses (anchoring bias), (2) excessive or inappropriate investigation ordering, (3) premature diagnostic closure, and (4) failure to consider or rule out critical, time-sensitive conditions [35, 27, 10].
- Our work has significant implications for medical AI development, providing a tool for assessing conversational AI systems intended for real-world clinical decision support. Beyond medicine, we

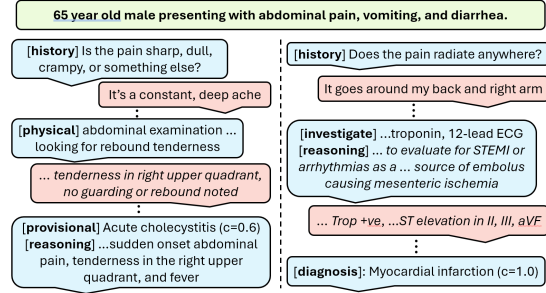


Figure 1: Action and reasoning trace of two evaluated models on our simulated *viva voce* examination. Given the initial scenario (green), agents (blue) are tasked to diagnose the patient (orange), which was simulated by our evaluation framework. c indicates confidence of diagnosis. Failure to perform a targeted clinical review (left) could lead to significant ramifications, such as a missed diagnosis of heart attack (correctly diagnosed on the right).

contribute to a growing body of work on agentic AI by demonstrating how sequential reasoning trajectories can diverge in complex, information-gathering decision-making environments.

2 Related Work

Early evaluations for language models in medicine, such PubMedQA [25] and MedQA [24], were based on static question-answering (QA) and assessed knowledge recall from complete information. While valuable for gauging basic clinical knowledge [43], these single-turn evaluations inadequately capture the dynamic, iterative reasoning essential for clinical practice [30]. High accuracy on such tasks can mask crucial metacognitive deficiencies, such as overconfidence or failure to recognize knowledge gaps, and may not translate to effective real-world decision-making, where LLMs have shown limitations even with full information [22, 35, 10]. Recent work like MetaMedQA has begun to explore these metacognitive aspects, revealing significant shortcomings in LLMs’ ability to handle underspecification and uncertainty both inside and outside medical contexts [39, 20].

Several other multi-turn benchmarks have also emerged to further assess sequential reasoning in medical LLMs. In particular, AI Hospital [14] and AgentClinic [40] simulate clinical encounters where LLMs must actively gather information. Both benchmarks showed that model performance degrades significantly in multi-turn settings compared to single-turn QA [14, 28, 29], suffering from error propagation and difficulties in maintaining context over extended interactions [21, 44]. Concurrently, research into conversational diagnostic AI, such as Google’s AMIE [47], included sophisticated benchmarks with multi-turn medical dialogue. While these benchmarks advance evaluation by incorporating interactivity, they rely on either a human-in-the-loop, or multiple LLM calls within their evaluation chain to account for different agent roles, potentially introducing non-determinism or data leakage, and may not be fully open-source or easily extensible.

Table 1: Comparison of Medical AI Datasets and Frameworks

Dataset/Framework	Evaluated Capabilities	Multi-turn	Det. Output	Open Source	Det. Criteria
QA Datasets [25, 24]	Knowledge	×	—	✓	✓
AI Hospital [14]	Diagnosis	✓	×	✓	×
AgentClinic [40]	Diagnosis, Bias	✓	×	✓	×
Hager et al. [22]	Decision making and Diagnosis	✓	✓	×	✓
Med-PaLM [43]	Diagnosis, Conversation Quality	✓	✓	×	✓
VivaBench (ours)	Diagnosis, Bayesian reasoning, Information Extraction	✓	✓	✓	✓

✓: Yes ×: No —: Not specified or N/A.

3 VivaBench

VivaBench simulates authentic clinical encounters where information must be actively gathered rather than being provided upfront, mirroring real-world diagnostic processes. We implement this by transforming free-text clinical data into a formalized Clinical Case ($\mathcal{C}$) structure. Each Case consists of five primary components that align with standard clinical documentation: History ($\mathcal{H}$), Physical Examination ($\mathcal{P}$), Imaging ($\mathcal{I}$), Laboratory investigations ($\mathcal{L}$), and a ground truth Diagnosis set ($\mathcal{D}$) that includes accepted differential diagnoses ($\mathcal{D}'$).

During evaluation, an agent $\mathcal{A}$ receives an initial clinical stem containing limited background information, and is tasked to progressively work toward a diagnosis with appropriate confidence. The agent can access four distinct action categories corresponding to the structured information components ($\mathcal{H}$, $\mathcal{P}$, $\mathcal{I}$, $\mathcal{L}$). Following each action, an examiner module ($\mathcal{E}$) retrieves and presents the specifically requested clinical information, effectively simulating the progressive information exchange characteristic of clinical viva voce examinations.

The evaluation proceeds through two distinct phases: (i) a **Review** phase, where the agent interviews the patient ($\mathcal{H}$) and conducts physical examination ($\mathcal{P}$), and (ii) an **Investigation** phase, where the

agent orders laboratory tests or imaging studies ($\mathcal{I}$, $\mathcal{L}$) to refine diagnostic hypotheses. The agent provides a **provisional diagnosis** with associated confidence level after the **Review** phase, followed by a **final diagnosis** after the Investigation phase, at which point the interaction concludes. Both diagnostic assessments are evaluated against the ground truth diagnosis set.

3.1 Dataset Creation

Dataset Source and Filtering A foundational objective of VivaBench is to create an open, accessible benchmark. As such, we sourced clinical vignettes exclusively from publicly available repositories, including MedQA [24], training materials from physician colleges in Australia and the United Kingdom [34], and PubMed case reports [16]. We selected cases based on predefined criteria optimizing for clinical relevance, diagnostic complexity, and educational value. Our selection prioritized cases that: (1) represent common clinical presentations, (2) necessitate multifaceted reasoning across different information types, (3) follow clear diagnostic pathways, and (4) include conditions with significant consequences if misdiagnosed. Appendix B provides a comprehensive description of our selection criteria.

Structuring Clinical Data To parse any free-text clinical information into a structure format, a standardized format was required. We designed an ontology according to standard clinical documentation guidelines [5, 1], and implemented a hierarchical schema for organizing clinical information. Each Clinical Case ($\mathcal{C}$) is partitioned into History ($\mathcal{H}$), Physical Examination ($\mathcal{P}$), Investigations ($\mathcal{I}$), and Laboratory tests ($\mathcal{L}$). History items are further characterized through attributes corresponding to the SOCRATES framework (Site, Onset, Character, Radiation, Associated symptoms, Time course, Exacerbating/relieving factors, Severity) for pain and general symptoms [19], supplemented by standard history categories (past medical history, medications, allergies, family history, social history). Physical examination findings are systematically organized by body systems. Investigations and Imaging are categorized by modality and specimen type, with each diagnostic test result accompanied by appropriate reference ranges where applicable.

To ensure standardization and facilitate information retrieval, each clinical data element is mapped to established medical terminologies (SNOMED-CT for $\mathcal{H}$ and $\mathcal{P}$, LOINC codes for $\mathcal{I}$ and $\mathcal{L}$, and ICD-10 codes for $\mathcal{D}$). This standardization enables precise matching between agent queries and available information. Appendix A provides a comprehensive description of our dataset schema.

Generation Pipeline Our dataset generation pipeline consists of both automated processing steps to handle volume and human review steps to ensure quality. We first screened for cases in large databases that fit our selection criteria with traditional NLP techniques and LLM assistance. Then, clinical information was extracted from source materials and systematically categorized into respective information components ($\mathcal{H}$, $\mathcal{P}$, $\mathcal{I}$, $\mathcal{L}$). Each component was further subdivided and assigned retrieval keys. Diagnoses were either parsed directly from the original vignettes, or annotated by clinical experts when not explicitly stated. In cases where information was implicit or incomplete, clinicians provided appropriate additions based on standard clinical knowledge to ensure comprehensive case representation. Finally, all structured clinical vignettes underwent physician review to verify information accuracy before inclusion in the dataset. We have made our generation codebase publicly available to enable further case generation from any clinical source material.

Dataset Description Our generation pipeline produced 1,952 cases of parsed structured data. Following human review, 990 cases were retained, all derived from the PubMed database [16]. Both the complete generated dataset and human-reviewed subset are available on our HuggingFace repository. Each entry includes a unique identifier (*uid*), *source* information, free-text clinical *vignette*, *diagnosis* with *differentials*, and a *clinicalcase* field containing a JSON string representing a Case object compatible with our evaluation framework. Appendix B provides detailed documentation of our dataset generation methodology.

3.2 Evaluation framework

Our evaluation framework provides the code infrastructure to evaluate any conversational AI agent $\mathcal{A}$ against our structured clinical cases. A unit of evaluation in our framework is an Examination, with an Examiner module ($\mathcal{E}$). $\mathcal{E}$ is responsible for the majority of work in our evaluation, including processing information requests, retrieving relevant data from the structured case, and returning information in a natural language format to $\mathcal{A}$. The evaluation workflow proceeds as follows:

Algorithm 1 Clinical Diagnostic Evaluation Process

- 1: $\mathcal{A}$ is presented with an initial clinical stem, and enters the **Review** phase of evaluation.
 - 2: $\mathcal{A}$ queries for bedside information using natural language
 - 3: $\mathcal{E}$ processes these queries, and returns relevant information from $\mathcal{H}$ and $\mathcal{P}$
 - 4: Steps 2, 3 are repeated, until $\mathcal{A}$ submits a provisional diagnosis, and proceeds to the **Investigation** phase of evaluation.
 - 5: $\mathcal{A}$ orders diagnostic investigations, including both imaging and laboratory investigations.
 - 6: $\mathcal{E}$ processes these queries, and returns relevant information from $\mathcal{I}$ and $\mathcal{L}$
 - 7: Steps 5, 6 are repeated, until $\mathcal{A}$ submits a final diagnosis
 - 8: Agent performance is evaluated based on accuracy of provisional and final diagnosis.
-

Information Retrieval and Parsing A core technical challenge in our framework is the translation between natural language queries and structured information retrieval. This process involves four sequential steps: receiving the free-text query from the agent, mapping this query to structured keys, retrieving structured information based on matched keys, and transforming this structured information into a natural language response. We provide both deterministic and LLM-based implementations of the mapper and parser components. The deterministic mapper employs cosine similarity embeddings to identify the most relevant keys, complemented by keyword matching and medical entity recognition. In contrast, the LLM-based mapper utilizes few-shot examples to interpret the semantic intent of queries and match them to appropriate information categories and keys. For history and physical examination findings, negative results (absent symptoms or normal examination findings) are explicitly returned when queried. Standardized laboratory values not specifically mentioned in the case are returned as default normal values with appropriate reference ranges, while investigations not available in the case are explicitly noted as “not available” to prevent information leakage. To validate our information retrieval system, we measured inter-annotator agreement on query-to-information mapping. We outline our experiments and validation results in the Appendix.

Metrics To evaluate diagnostic performance, we assessed both accuracy and diagnostic reasoning process. We computed top-k exact and approximate accuracies for $k=1$ through 5, where a prediction was considered an exact match if its ICD-10 code matched a ground truth diagnosis at the appropriate hierarchical level (e.g., a prediction of E78.1 matches ground truth E78), or an approximate match if it matched an accepted differential, shared the top levels of ICD-10 code with any diagnosis, or had high semantic similarity to the correct diagnosis. For cases with multiple ground truth diagnoses, accuracy was counted if any one diagnosis was correctly identified. We also measured model confidence using a weighted score:

$$S_{\text{conf}} = \sum_{i \in E} c_i + \sum_{i \in A} c_i - \sum_{i \in U} c_i$$

where c_i represents the normalized confidence for each prediction belonging to exact matches (E), approximate matches (A), or unmatched predictions (U). This score ranges from -1.0 (high confidence in wrong diagnoses) to 1.0 (high confidence in correct diagnoses).

Additionally, we tracked the relevance of clinical information requested by calculating precision (proportion of requested information that was relevant to the diagnosis: $|K_{\text{requested}} \cap K_{\text{relevant}}|/|K_{\text{requested}}|$) and recall (proportion of all diagnosis-relevant information that was requested: $|K_{\text{requested}} \cap K_{\text{relevant}}|/|K_{\text{relevant}}|$). These metrics were calculated for different information categories (history/physical vs. investigations) and scopes (targeted to matched diagnoses vs. all potential diagnoses). Finally, we measured how diagnoses evolved from provisional to final stages to evaluate how effectively models updated their reasoning when presented with additional information.

4 Evaluation

We evaluated several state-of-the-art foundation models, including Gemini 2.5 pro [18], DeepSeek-R1 [13], o4-mini from OpenAI [36], Llama-4 Maverick [33], Grok 3 mini beta [48], and Qwen 3 (235b-a22b) [8]. Each model was tested at temperature 0, and instructed with the task to perform a clinical review and diagnosis on the patient, with clear instructions and explanation on the Review and Investigation phases. The full prompt and details of our evaluation task are outlined in Appendix E.

Table 2: Comparative performance of evaluated LLMs on VivaBench.

Model	Actions			Top-k P. (d_P)			Top-k D. (d_D)			Top-k F. (d_F)			Raw Conf.		S_{conf}	
	Total	Rv	Ix	1	3	5	1	3	5	1	3	5	C_P	C_D	S_P	S_D
Gemini 2.5 Pro	8.8	4.4	2.3	0.17	0.30	0.33	0.35	0.46	0.48	0.69	0.84	0.86	0.45	0.77	0.25	0.37
DeepSeek-R1	5.5	2.0	1.5	0.12	0.22	0.24	0.23	0.31	0.32	0.61	0.77	0.80	0.33	0.59	0.22	0.35
o4-mini	8.9	4.0	2.9	0.15	0.24	0.26	0.32	0.40	0.41	0.63	0.79	0.81	0.38	0.78	0.25	0.44
Llama-4 Maverick	8.5	3.0	3.5	0.08	0.17	0.18	0.23	0.27	0.27	0.52	0.70	0.71	0.49	0.70	0.20	0.37
Grok 3 Mini Beta	7.0	3.4	1.6	0.10	0.20	0.23	0.16	0.27	0.29	0.60	0.76	0.77	0.30	0.36	0.20	0.26
Qwen 3	5.5	1.8	1.8	0.10	0.19	0.21	0.21	0.30	0.31	0.47	0.66	0.71	0.29	0.47	0.20	0.29

Actions: Average interaction turns per case (Total), broken down by Review queries ($Rv = \mathcal{H} + \mathcal{P}$) and Investigation ($Ix = \mathcal{I} + \mathcal{L}$) orders. **Top-k:** Top-k Exact Accuracy for Provisional (d_P) (**P**), Final (d_D) (**F**) (after interactive information gathering), and Full Information (d_F) (**F**) Diagnoses at $k \in \{1, 3, 5\}$. C_P, C_D : Mean raw confidence scores for provisional and final diagnoses, respectively. S_P, S_D : Confidence-Weighted Accuracy Scores (S_{conf}) for provisional and final diagnoses, respectively.

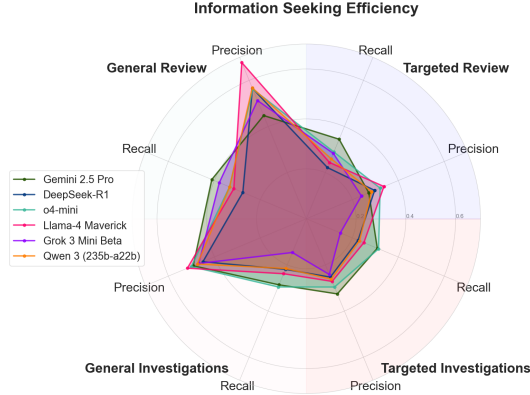


Figure 2: Radar plot comparing precision and recall metrics for how effectively models gather clinical information. **Targeted** metrics assess performance on gathering diagnosis-relevant information only, while **Overall** metrics include all available clinical information. **Review** includes history-taking and physical examination, while **Investigations** covers labs and imaging. Higher values are better.

4.1 Results

Diagnostic Accuracy The diagnostic performance results are outlined in Table 2. Overall, while most models demonstrated competence in deriving diagnoses when presented with complete clinical information (Full Information condition), they struggled significantly when constrained to the interactive examination format. Gemini 2.5 Pro consistently outperformed other models across all diagnostic stages and metrics, achieving the highest top-1 accuracy in final diagnosis (35%) and full information scenarios (69%). This performance difference can be attributed with the relative recency and scale of the models evaluated, suggesting that larger, more recent models may better handle the complex reasoning required for sequential diagnostic tasks. Nonetheless, there is a substantial performance gap between final diagnosis accuracy and full information accuracy across all models, with most models at least doubling in accuracy performance. This suggests that models possess the requisite knowledge to come to a diagnosis (or straightup has the data in the training set), but they struggle with handling uncertainty or to undergo the targeted information-seeking process needed to apply it effectively. The progression from top-1 to top-5 accuracy (Figure 3) shows diminishing returns across all models, with most significant gains occurring between top-1 and top-3. This suggests that when models are incorrect in their primary diagnosis, the correct answer is often present within their top alternatives, indicating partial understanding of the clinical picture.

Subgroup analysis demonstrated variation in model performance across different specialty groups (Figure 4). All models demonstrated relative strengths in Infectious Disease & Immunology and Cardiovascular & Metabolic conditions, likely reflecting the prevalence of these conditions in medical literature and their often distinctive presentation patterns. Conversely, most models showed comparative weakness in Pediatric and Neurological/Psychiatric conditions, which frequently require more nuanced interpretation of symptoms and developmental context.

Information Seeking Efficiency To evaluate how models gather clinical information strategically, we measured their precision and recall across different information-seeking activities. We define "information seeking efficiency" as the ability to selectively request relevant history, physical exami-

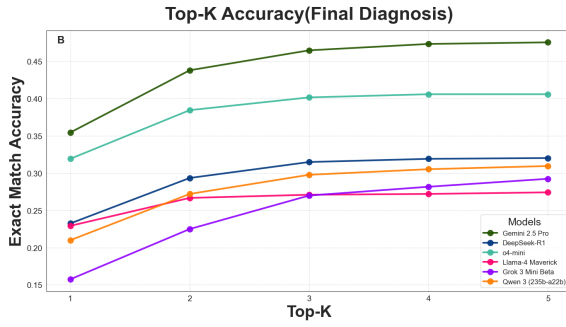


Figure 3: Accuracy of top-1 to top-5 diagnoses.

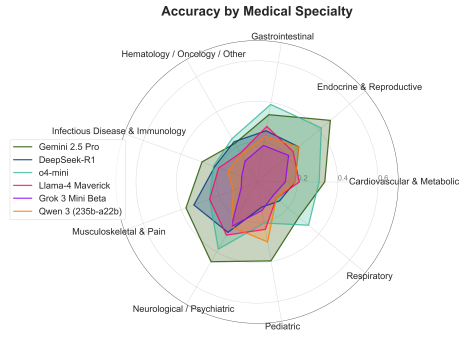


Figure 4: Performance across specialties.

Table 3: Correlation analysis of diagnostic adaptation behaviors. The table shows Pearson correlation coefficients between diagnostic adaptation **Variables** (center-right columns) and performance **Outcomes** (left data columns). **Outcomes** are defined as the change from provisional (d_p) to final (d_f) stage in: Top-1 Exact Accuracy (Acc Δ), Top-1 Approximate Accuracy (Apr. Acc Δ), and Confidence-Weighted Score (S_{conf} Δ). **Variables** include: counts of Diagnoses Added, Removed, or Maintained; Confidence Delta (overall change in mean raw confidence, $C_D - C_P$); Confidence Shift (mean $c_f - c_p$) and Confidence Shift Magnitude (mean $|c_f - c_p|$) for maintained diagnoses. Superscripts indicate statistically significant correlations ($p < 0.05$): $\ddagger$ with Exact Accuracy Change; † with Approximate Accuracy Change; * with S_{conf} Change.

Model	Outcomes (Change $d_p \rightarrow d_f$)			Diagnostic Adaptation Variables					
	Acc Δ	Apr. Acc Δ	S_{conf} Δ	Diag. Added	Diag. Removed	Diag. Maintained	Conf. Delta	Conf. Shift	Conf. Shift Mag.
Gemini 2.5 Pro	0.18	0.09	0.12	2.25 *	3.08 $^\dagger\ddagger^*$	0.76 $^\dagger\ddagger^*$	0.32 $^\dagger\ddagger^*$	0.10	0.19 *
DeepSeek-R1	0.11	0.08	0.13	1.69 $^\dagger\ddagger^*$	2.65 $^\dagger\ddagger^*$	0.90 $^\dagger\ddagger^*$	0.26 $^\dagger\ddagger^*$	0.12 $^\dagger\ddagger^*$	0.24
o4-mini	0.17	0.10	0.18	1.24 ‡	2.47 $^\dagger\ddagger^*$	0.70 $^\dagger\ddagger^*$	0.40 $^\dagger\ddagger^*$	0.12	0.20 *
Llama-4 Maverick	0.14	0.10	0.17	1.11	1.79 $^\dagger\ddagger^*$	0.71 $^\dagger\ddagger^*$	0.21 $^\dagger\ddagger^*$	0.05 $^\dagger\ddagger^*$	0.17
Grok 3 Mini Beta	0.05	0.03	0.06	1.40 ‡	1.64 $^\dagger\ddagger^*$	2.70 $^\dagger\ddagger^*$	0.05 $^\dagger\ddagger^*$	0.00 ‡	0.42 ‡
Qwen 3 (235b-a22b)	0.11	0.06	0.08	1.94 †‡	2.55 $^\dagger\ddagger^*$	1.34 *	0.18 $^\dagger\ddagger^*$	0.07 ‡	0.27 *

nation findings, and diagnostic investigations that contribute to accurate diagnosis. This provides insight into the diagnostic strategies models employ—precision reflects how targeted their inquiries are toward specific diagnoses, while recall indicates their thoroughness in confirming diagnoses and excluding alternatives.

Figure 2 shows distinct information-seeking patterns across models. Gemini 2.5 Pro demonstrated the most balanced approach, with strong performance in both overall information gathering (breadth) and diagnosis-targeted inquiries (depth). This balanced information-seeking strategy likely contributed to its superior diagnostic accuracy. Most models showed asymmetric patterns, with generally higher precision than recall across all categories, suggesting they are selective in their inquiries but often miss relevant clinical details.

Models generally performed better at selecting appropriate clinical review items than ordering relevant investigations. The low recall in diagnostic investigations could be attributed to the abundance of clinical information within the vignettes. However, their precision were also low in terms of ordering the exact tests that would help them with diagnoses. While a lot of the requested tests were basic routine blood tests that were not included in the vignette, we also observe that models were ordering a lot of random and super niche stuff, such as genetic testing. It is likely that this was due to training, where models would naturally be trained on the more specific / niche stuff in an academic setting.

Diagnostic Adaptation To further quantify how models adjust their decision making process in light of uncertainty, we measured their diagnostic adaptation behaviors and performance outcomes. In particular, we measured changes in diagnoses between provisional and final stages, and also how the model’s confidence distribution changed between steps. We tracked absolute confidence changes, defined as the difference averaged value of confidence of diagnoses between stages, and the shift of confidence in maintained diagnoses (diagnoses sustained between provisional and final diagnosis stage). Table 3 shows correlations between diagnostic adaptation behaviors and performance outcomes, as measured by the delta change in exact and top-1 accuracy, as well as confidence S.

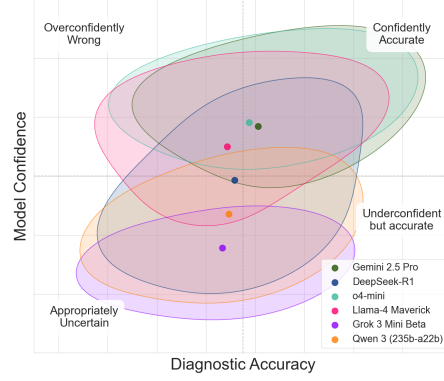


Figure 5: Kernel density estimation (KDE) contours representing the distribution model performance across clinical cases. The quadrants reflect differences in confidence and accuracy of diagnosis: confidently accurate (top-right), overconfident and wrong (top-left), underconfident but accurate (bottom-right), and appropriately uncertain (bottom-left).

Diagnosis removal and diagnosis maintenance both showed significant correlation to all 3 outcome metrics. This is expected, as models that correctly identified diagnoses during the provisional stage only needed to further investigate and rule out their other suspected conditions. In contrast, diagnosis addition showed less significant correlation to outcomes, representing a more challenging adaptation task. We observed divergent patterns between model performance tiers: lower-performing models showed improved accuracy but not confidence when adding diagnoses, suggesting more random but less confident guesses. Better-performing models showed improvements in both metrics or primarily in confidence scores, indicating they could recognize when initial diagnostic considerations were incomplete and adjust accordingly.

Confidence A consistent pattern across all models was the strong positive correlation between absolute confidence changes and diagnostic performance improvements. As models gathered more information, those that demonstrated larger shifts in overall confidence (regardless of direction) tended to show greater improvements in both exact and approximate diagnostic accuracy. This suggests that meaningful engagement with new clinical information is reflected in confidence recalibration and serves as an indicator of effective diagnostic reasoning.

All models tended to increase confidence in diagnoses they maintained between stages, regardless of whether those diagnoses were correct. The magnitude of confidence change for maintained diagnoses correlated strongly with overall confidence score changes, but showed limited correlation with accuracy improvements. This suggests that models' growing certainty often developed independently of diagnostic correctness, indicating a form of confirmation bias.

DeepSeek and Llama displayed significant correlations between the direction of confidence shifts and accuracy changes, suggesting more calibrated confidence adjustments, where they increase confidence primarily when moving toward correct diagnoses. This behavior appears related to their information-seeking efficiency, as both models demonstrated higher precision in their review and diagnostic metrics, suggesting that their ability to seek targeted information allows them to hone in on the correct diagnosis, while increasing their confidence in doing so.

To further explore the relationship between diagnostic performance and confidence calibration, we mapped models' confidence against accuracy using KDE distributions (Figure 5). This analysis revealed distinct calibration across model architectures. Gemini 2.5 Pro and o4-mini demonstrated well-calibrated confidence, positioning primarily in the "confidently accurate" quadrant. DeepSeek-R1 exhibited more conservative confidence estimation despite reasonable accuracy, while Llama-4 Maverick showed a more complex pattern with higher overall confidence despite mixed accuracy results. Grok 3 Mini Beta and Qwen 3 displayed lower confidence profiles, and had a diverse range of accuracy metrics across the full dataset. These distinctions highlight fundamental differences in how models internally calibrate uncertainty during diagnostic reasoning, with important implications for clinical deployment where both accurate diagnoses and appropriate confidence signaling are essential.

4.2 Failure Mode Analysis

To systematically characterize why agents fail in clinical reasoning, we conducted both qualitative review of individual reasoning traces and quantitative classification across the full evaluation set. Through initial review of agent interactions, we identified three types of failures that mirror well-documented cognitive errors in clinical practice: (1) *inappropriate hypothesis generation*, where agents prioritize clinically implausible diagnoses given patient demographics or epidemiology; (2) *premature diagnostic closure* (satisfaction-of-search bias), where agents terminate investigation after confirming an initial hypothesis while missing causative or concurrent conditions; and (3) *inadequate investigations*, where agents order diagnostically insufficient tests or fail to account for test limitations given the clinical presentation.

Quantitative Analysis To quantify the prevalence of these failure modes, we employed an LLM-based classifier (GPT-4.1) to categorize 738 evaluation traces from failed diagnostic attempts. The classifier used few-shot examples of each failure mode to systematically label agent reasoning traces. The distribution of failures was: inappropriate hypothesis generation (348 cases, 47%), premature diagnostic closure (291 cases, 39%), inadequate investigations (90 cases, 12%), and ICD-10 coding errors (9 cases, 1%). These findings reveal that the majority of model failures stem from reasoning deficits—specifically anchoring bias and satisfaction-of-search bias—rather than knowledge gaps. Limitations in clinical knowledge (e.g., selecting appropriate diagnostic investigations) accounted for a smaller proportion of errors.

Qualitative Analysis To provide concrete illustrations of these failure patterns, we present detailed case studies in Appendix G that demonstrate how each failure mode manifests in agent reasoning trajectories. These examples highlight scenarios that could compromise patient safety through missed diagnoses, delayed interventions, or inappropriate clinical pathways if current LLM agents were deployed in real-world clinical settings. Notably, the observed failure modes parallel common cognitive biases documented in human clinical reasoning [35, 27, 10], suggesting that current LLMs may replicate rather than mitigate well-known sources of diagnostic error.

5 Discussion

Our evaluation reveals a significant gap between LLMs’ knowledge base and their ability to engage in sequential clinical reasoning. The performance disparity between full-information and interactive examination conditions demonstrates that current models possess relevant medical knowledge, but struggle with systematic information gathering and hypothesis refinement under uncertainty. This leads to three patterns of failure modes, with inefficient information seeking (high precision but low recall), suboptimal hypothesis revision (difficulty adding appropriate new diagnoses), and confidence miscalibration (increasing confidence regardless of diagnostic correctness). These findings highlight how current LLMs, despite superior knowledge retrieval capabilities, exhibit reasoning patterns that diverge from sound clinical judgment when navigating diagnostic uncertainty. VivaBench connects to several active research areas including sequential decision-making, reasoning agents, confidence calibration, and strategic information gathering, and provides a structured environment for the wider machine learning community to study how models make decisions with partial information.

Limitations Our study has several limitations. The dataset derives from a single source type (clinical case reports), and remains modest in scale compared to large-scale medical benchmarks, due to the labour intensity of the data generation process. The deterministic variant of our information retrieval system does not capture the full variability in real clinical communication, while our LLM variant is not fully deterministic, affecting the validity of our benchmark. Additionally, due to computational constraints, we conducted only a single evaluation run for each model, which may not account for the stochastic nature of LLM outputs, especially over long horizons. Lastly, despite structuring cases according to clinical standards, the *viva voce* examination format remains a simplified approximation of actual patient encounters, lacking the multi-faceted nature and nuance in clinical practice.

References

- [1] Clinical Practice Guideline Manual. URL <https://www.aafp.org/family-physician/patient-care/clinical-recommendations/cpg-manual.html>.
- [2] ICD-10 Version:2019. URL <https://icd.who.int/browse10/2019/en>.
- [3] LOINC. URL <https://loinc.org/>.
- [4] OpenRouter. URL <https://openrouter.ai>.
- [5] RACGP - Clinical guidelines. URL <https://www.racgp.org.au/clinical-resources/clinical-guidelines>.
- [6] SNOMED CT. URL <https://www.nlm.nih.gov/healthit/snomedct/index.html>.
- [7] A. I. H. Abuzied and W. O. M. Nabag. Structured viva validity, reliability, and acceptability as an assessment tool in health professions education: a systematic review and meta-analysis. *BMC Medical Education*, 23(1):531, 2023.
- [8] Alibaba Cloud. Qwen3-235b-a22b, April 2025. URL <https://qwen.ai/blog/qwen3>. Accessed: May 16, 2025. Model card also available on Hugging Face. [24].
- [9] Artificial Analysis. LLM Leaderboard - Comparison of GPT-4o, Llama 3, Mistral, Gemini and over 30 models. <https://artificialanalysis.ai/leaderboards/models>, 2024. Accessed: 2025-05-23.
- [10] F. Busch, L. Hoffmann, C. Rueger, E. H. van Dijk, R. Kader, E. Ortiz-Prado, M. R. Makowski, L. Saba, M. Hadamitzky, J. N. Kather, et al. Systematic review of large language models for patient care: current applications and challenges. *MedRxiv*, pages 2024–03, 2024.
- [11] F. Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.
- [12] M. Daniel, J. Rencic, S. J. Durning, E. Holmboe, S. A. Santen, V. Lang, T. Ratcliffe, D. Gordon, B. Heist, S. Lubarsky, et al. Clinical reasoning assessment methods: a scoping review and practical guidance. *Academic Medicine*, 94(6):902–912, 2019.
- [13] DeepSeek AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL <https://arxiv.org/abs/2501.12948>.
- [14] Z. Fan, J. Tang, W. Chen, S. Wang, Z. Wei, J. Xi, F. Huang, and J. Zhou. Ai hospital: Interactive evaluation and collaboration of llms as intern doctors for clinical diagnosis. *arXiv e-prints*, pages arXiv–2402, 2024.
- [15] J. Fashner and A. C. Gitu. Diagnosis and treatment of peptic ulcer disease and h. pylori infection. *American family physician*, 91(4):236–242, 2015.
- [16] T. N. C. for Biotechnology Information. Pubmed, 2022. URL <https://huggingface.co/datasets/ncbi/pubmed>.
- [17] S. Golchin and M. Surdeanu. Time Travel in LLMs: Tracing Data Contamination in Large Language Models, Feb. 2024.
- [18] Google. Gemini 2.5 pro, 2025. URL https://vertexaisearch.cloud.google.com/grounding-api-redirect/AbF9wXGhnLIvwm057uhI8uEPCY6jAn_ESE6kjSe566rsfiPnZ0V86ieMAM_IE61sNHfIiVUsqoIJADnsGaELwr5Wix9JAUe9tXTErNAkvob9DJeo06ZWjRM6_rTQF0bowTortUT1Pxhx6BwQVdgi1BC1D0RgwCeLrLhmXT-dtKw4PFHGQ==. Accessed: May 16, 2025.
- [19] J. Gregory. Use of pain scales and observational pain assessment tools in hospital settings. *Nurs Stand*, 34(9):70–4, 2019.
- [20] M. Griot, C. Hemptinne, J. Vanderdonckt, and D. Yuksel. Large language models lack essential metacognition for reliable medical reasoning. *Nature communications*, 16(1):642, 2025.
- [21] S. Guan, H. Xiong, J. Wang, J. Bian, B. Zhu, and J.-g. Lou. Evaluating llm-based agents for multi-turn conversations: A survey. *arXiv preprint arXiv:2503.22458*, 2025.
- [22] P. Hager, F. Jungmann, R. Holland, K. Bhagat, I. Hubrecht, M. Knauer, J. Vielhauer, M. Makowski, R. Braren, G. Kaissis, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622, 2024.

- [23] D. Y. Hwang, G. S. Silva, K. L. Furie, and D. M. Greer. Comparative Sensitivity of Computed Tomography vs. Magnetic Resonance Imaging for Detecting Acute Posterior Fossa Infarct. *The Journal of Emergency Medicine*, 42(5):559–565, 2012. ISSN 0736-4679. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3346849/>.
- [24] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.
- [25] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019.
- [26] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1): 1–9, 2016.
- [27] J. Kim, B. N. Vajravelu, et al. Assessing the current limitations of large language models in advancing health care education. *JMIR Formative Research*, 9(1):e51319, 2025.
- [28] W.-C. Kwan, X. Zeng, Y. Jiang, Y. Wang, L. Li, L. Shang, X. Jiang, Q. Liu, and K.-F. Wong. Mt-eval: A multi-turn capabilities evaluation benchmark for large language models. *arXiv preprint arXiv:2401.16745*, 2024.
- [29] P. Laban, H. Hayashi, Y. Zhou, and J. Neville. LLMs get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*, 2025.
- [30] F. Lateef. Clinical reasoning: The core of medical education and practice. *European Journal of Cardiovascular Medicine*, 11(3), 2021.
- [31] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.
- [32] C.-Y. Lin. ROUGE: A Package for Automatic Evaluation of Summaries. *ACL Anthology*, 2004.
- [33] Meta. Llama 4 maverick, April 2025. URL https://vertexaisearch.cloud.google.com/grounding-api-redirect/AbF9wXHktpFbd1Z7LI_cPPrrRgtFEU1NsvCWbtIdqEFFmApPqEm-aqb_UGNAjJzdGG9ZRxtWdTGdk_1t4zVPOUxb9LDfh0iP11Ay9CqgleKbWxgreBAnWzrgubkttkMXTzXbeqWIh69IA5i9-eX11R1ISHu8wsCageBBXpjJd0=. Accessed: May 16, 2025. Model card available on Hugging Face and other platforms.
- [34] T. F. of the Royal Colleges of Physicians. Sample scenarios. URL <https://www.thefederation.uk/examinations/paces/sample-scenarios>.
- [35] J. A. Omiye, H. Gui, S. J. Rezaei, J. Zou, and R. Daneshjou. Large language models in medicine: the potentials and pitfalls: a narrative review. *Annals of internal medicine*, 177(2):210–220, 2024.
- [36] OpenAI. Openai o3 and o4-mini system card, April 2025. URL <https://openai.com/index/o3-o4-mini-system-card/>. Accessed: May 16, 2025.
- [37] L. Phan, A. Gatti, Z. Han, N. Li, J. Hu, H. Zhang, C. B. C. Zhang, M. Shaaban, J. Ling, S. Shi, et al. Humanity’s last exam. *arXiv preprint arXiv:2501.14249*, 2025.
- [38] M. Renze and E. Guven. The Effect of Sampling Temperature on Problem Solving in Large Language Models, Feb. 2024.
- [39] Y. Ruan, H. Dong, A. Wang, S. Pitis, Y. Zhou, J. Ba, Y. Dubois, C. J. Maddison, and T. Hashimoto. Identifying the risks of lm agents with an lm-emulated sandbox. In *International Conference on Learning Representations*, 2024.
- [40] S. Schmidgall, R. Ziaei, C. Harris, E. Reis, J. Jopling, and M. Moor. AgentClinic: A multimodal agent benchmark to evaluate AI in simulated clinical environments, 2024. URL <http://arxiv.org/abs/2405.07960>.
- [41] T. Schubert, R. W. Peck, A. Gimson, C. Davtyan, and M. van der Schaar. A foundational framework and methodology for personalized early and timely diagnosis. *arXiv preprint arXiv:2311.16195*, 2023.
- [42] T. Sellam, D. Das, and A. P. Parikh. BLEURT: Learning Robust Metrics for Text Generation, May 2020.
- [43] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

- [44] V. Sirdeshmukh, K. Deshpande, J. Mols, L. Jin, E.-Y. Cardona, D. Lee, J. Kritz, W. Primack, S. Yue, and C. Xing. Multichallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier LLMs. *arXiv preprint arXiv:2501.17399*, 2025.
- [45] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shueb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.
- [46] L. N. Steimle and B. T. Denton. Markov decision processes for screening and treatment of chronic diseases. *Markov Decision Processes in Practice*, pages 189–222, 2017.
- [47] T. Tu, A. Palepu, M. Schaekermann, K. Saab, J. Freyberg, R. Tanno, A. Wang, B. Li, M. Amin, N. Tomasev, S. Azizi, K. Singhal, Y. Cheng, L. Hou, A. Webson, K. Kulkarni, S. S. Mahdavi, C. Semturs, J. Gottweis, J. Barral, K. Chou, G. S. Corrado, Y. Matias, A. Karthikesalingam, and V. Natarajan. Towards Conversational Diagnostic AI, 2024. URL <http://arxiv.org/abs/2401.05654>.
- [48] xAI. Grok 3 mini beta, February 2025. URL <https://x.ai/blog/grok-3>. Accessed: May 16, 2025. Some sources list an April 9, 2025 release for Grok 3 Mini. [30].
- [49] Z. Yi, J. Ouyang, Y. Liu, T. Liao, Z. Xu, and Y. Shen. A survey on recent advances in llm-based multi-turn dialogue systems. *arXiv preprint arXiv:2402.18013*, 2024.
- [50] C. Zhang, X. Dai, Y. Wu, Q. Yang, Y. Wang, R. Tang, and Y. Liu. A survey on multi-turn interaction capabilities of large language models. *arXiv preprint arXiv:2501.09959*, 2025.

A Schema Design

The base unit for each test case in the VivaBench benchmark is a Clinical Case ($\mathcal{C}$). Each Case is converted from a free-text clinical vignette leading up to a diagnosis, and comprises five primary components: History ($\mathcal{H}$), Physical Examination ($\mathcal{P}$), Imaging ($\mathcal{I}$), Laboratory investigations ($\mathcal{L}$), and a ground truth Diagnosis set ($\mathcal{D}$), which includes a list of accepted differential diagnoses ($\mathcal{D}'$).

Algorithm 2 Input schema for a structured Clinical Case $\mathcal{C}$

History ($\mathcal{H}$):

Demographics: Patient age, gender, ethnicity.

Chief Complaint: A one-line summary of the patient’s presentation.

Symptoms: A list of symptom objects, each with attributes including (but not limited to):

- **name:** Best matching SNOMED-CT description of the symptom.
- **onset:** When the symptom first began (e.g., "2 days ago").
- **duration:** How long the symptom has persisted (e.g., "3 hours").
- **progression:** How the symptom evolved over time.
- **timing:** When the symptom occurs (e.g., "morning", "after meals").
- **severity:** Intensity of the symptom (e.g., "mild", "severe").
- **system:** Body system affected (e.g., "cardiovascular").
- **location:** Anatomical location of the symptom.
- **character:** Quality or nature of the symptom (e.g., "sharp", "dull").
- **radiation:** Whether and where the symptom spreads.
- **alleviating_factors:** Factors that improve the symptom.
- **aggravating_factors:** Factors that worsen the symptom.
- **associated_symptoms:** Other co-occurring symptoms.
- **context:** Circumstances surrounding the symptom.
- **history:** Detailed narrative specific to the symptom, extracted from the input vignette.

Past Medical History: Previous diagnoses and conditions.

Medication History: Current medications with dosage and frequency.

Allergies: Known drug or food allergies.

Social History: Lifestyle factors (e.g., smoking, alcohol, occupation).

Family History: Relevant conditions in family members.

Physical Examination ($\mathcal{P}$): Vital signs and physical findings, organized by body systems.

Lab Investigations ($\mathcal{L}$): Laboratory and diagnostic test results, typically mapped to LOINC codes.

Imaging ($\mathcal{I}$): Radiological studies, categorized by modality and anatomical region.

Diagnosis ($\mathcal{D}$): A list of final diagnosis objects. Each object includes:

- The free-text label from the original vignette (e.g., "Tumefactive Crohn’s disease").
- The best-matching ICD-10 code and name (e.g., K50.8, "Crohn’s disease of both small and large intestine").
- A set of keys from $\mathcal{H}$, $\mathcal{P}$, $\mathcal{I}$, $\mathcal{L}$ that support this diagnosis.

Differentials ($\mathcal{D}'$): A list of alternative diagnosis objects. Each object includes:

- Its own ICD-10 code and name (if applicable).
 - A free-text field describing the clinical reasoning supporting it as a differential.
 - (Optionally, supporting keys as in $\mathcal{D}$).
-

Design considerations of clinical Schema. To convert free-text clinical vignettes into a machine-readable format suitable for VivaBench, we defined the five-part schema for a structured clinical case $\mathcal{C} = \{\mathcal{H}, \mathcal{P}, \mathcal{I}, \mathcal{L}, \mathcal{D}\}$. $\mathcal{H}$ (History) captures the patient’s clinical presentation and other relevant medical history. $\mathcal{P}$ (Physical Examination) records vital signs and examination findings, organized by body systems. $\mathcal{I}$ (Imaging) holds radiological and nuclear medicine studies, while $\mathcal{L}$ (Laboratory) contains all non-imaging investigations. $\mathcal{D}$ lists the final diagnosis (or diagnoses), and $\mathcal{D}'$ contains other accepted differential diagnoses. We refined this schema by reviewing standard clinical documentation guidelines [5, 1] and soliciting feedback from our clinician panel. Every field within the schema may be null or multi-valued, and corresponding excerpts from the input vignette are preserved for traceability to ensure fidelity to the source material.

Overview of Clinical Schema. Algorithm 2 describes our schema for $\mathcal{C}$ in detail. Within $\mathcal{H}$, the presenting complaint is expressed as a list of Symptom objects. Each symptom is characterized by

multiple attributes, with core attributes based on the SOCRATES framework [19] (e.g., Site, Onset, Character, Radiation, Associated symptoms, Time course, Exacerbating/relieving factors, Severity), supplemented by additional descriptive fields for more complex symptoms. When applicable, a time-stamped trajectory is included for any element that changes over time (e.g., "pain severity: mild $\rightarrow$ moderate $\rightarrow$ severe over 24 hours"). Relevant negative findings (e.g., "denies dyspnea") are also explicitly encoded to provide clinical context. $\mathcal{H}$ also includes distinct sub-sections for past medical and surgical history, current medications, allergies, social history, and family history.

All clinical concepts in $\mathcal{H}$ and $\mathcal{P}$ utilize SNOMED-CT codes [6] when available to ensure standardized terminology. Similarly, $\mathcal{L}$ and $\mathcal{I}$ items are mapped to LOINC procedural codes [3]. Final diagnoses ($\mathcal{D}$) and differentials ($\mathcal{D}'$) carry ICD-10 names and codes [2]. If a standard code is not available or applicable for any item, the original free-text label from the source vignette is used.

$\mathcal{D}$ can represent multiple final diagnoses if present in the source case. Each diagnosis entry specifies a condition name from the original vignette, its corresponding ICD-10 code and name, and the set of relevant keys from $\mathcal{H}$, $\mathcal{P}$, $\mathcal{I}$, and $\mathcal{L}$ that contribute to the diagnosis. A list of accepted differentials ($\mathcal{D}'$) is included for cases with inherent diagnostic ambiguity (e.g., when the original vignette considered multiple diagnoses) or those requiring specialized investigations (e.g., differentiating subtypes of lung cancer). In addition to their own ICD-10 mappings, each entry in $\mathcal{D}'$ includes a free-text field describing the clinical reasoning that supports it as a differential diagnosis.

Our schema aggregates data across these domains from a cross-sectional, first-visit perspective. Information regarding prior admissions or procedures not directly relevant to the patient’s current presentation is captured under past medical history within $\mathcal{H}$. Repeat admissions for the same presenting complaint are merged into the initial encounter by appending trajectories to the relevant data fields. For example, a patient who had open heart surgery two months prior and was discharged yesterday for stable angina, then readmitted with recurrent chest pain hours later, will have "open heart surgery 2 months ago" in the past medical history section of $\mathcal{H}$. All clinical information directly related to the current presentation (including the recurrent chest pain) is aggregated into their respective single fields within the current encounter, with temporal evolution preserved (e.g., symptoms: "Chest pain initially improved with GTN. However, re-emerged 24 hours later"; investigations: "Troponin: 5 $\rightarrow$ 7 $\rightarrow$ 11 ng/mL at 1h, 3h, 5h").

B Dataset Generation Pipeline

We employed a hybrid approach combining automated Natural Language Processing (NLP) heuristics, LLM-based screening, and human expert review to generate the VivaBench dataset. Figure 6 summarizes our dataset generation pipeline. The generation process begins with rules-based scoring over clinical case report databases to cases with comprehensive clinical information. The top scoring cases subsequently go through LLM-based screening against a set of pre-defined criteria to identify potentially relevant candidates. Human medical experts further review these candidates to ensure clinical accuracy and relevance. Selected clinical vignettes undergo transformation into our standardized schema detailed outlined in Appendix A. Lastly, generated clinical cases undergo another round of human review before inclusion to the final dataset.

Dataset Source To source clinical vignettes for VivaBench, we prioritized publicly accessible repositories providing free-text clinical data that included patient history, physical examination findings, investigation results, and clear diagnosis. Our exploration covered MedQA [24], training materials from physician colleges in Australia and the United Kingdom [34], and case reports from PubMed [16]. We also examined restricted-license patient datasets, such as MIMIC [26], during initial experiments to inform our schema design and data exploration strategies. Following a comprehensive review by our clinical team, PubMed case reports were selected as the primary source due to their wide variety, high quality of clinical detail, and rich diagnostic narratives spanning diverse medical specialties.

Screening and Filtering We implemented a two-stage filtering pipeline to identify the most educationally valuable cases from an initial pool of over 250,000 PubMed case reports. The first stage involved an initial screening pass using NLP techniques to analyze case report titles and content. This step aimed to identify reports containing substantive clinical review components and a clear diagnostic focus, while filtering out non-clinical reports, non-human subject studies, reports primarily

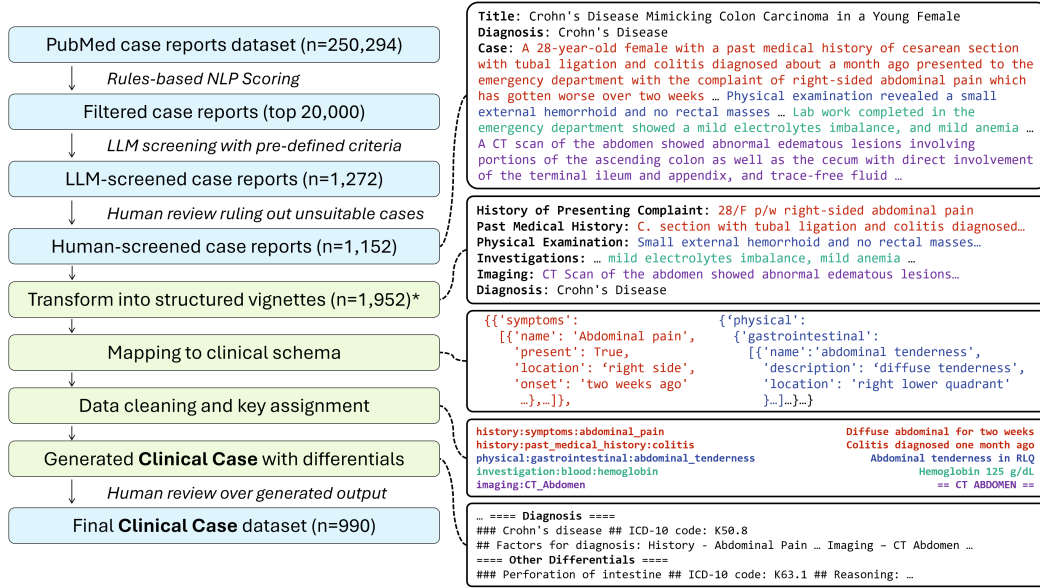


Figure 6: Our clinical case dataset generation pipeline, transforming PubMed case reports into structured, machine-learning-ready clinical vignettes for VivaBench. The workflow (left panel) illustrates filtering stages (blue boxes) that reduce an initial pool of 250,294 raw PubMed reports to 1,152 human-screened candidate cases, and subsequent transformation steps (green boxes) that result in 990 final structured clinical cases (*Additional cases from MedQA dataset were also generated in our initial experiments). The right panel provides an example of this transformation for a 28-year-old female with Crohn’s Disease, showing the conversion of unstructured medical text into a standardized case format. This structured representation preserves the clinical reasoning process essential for our evaluation framework.

describing surgical techniques, or those detailing exceptionally rare diseases. This rule-based filtering yielded approximately 20,000 candidate cases.

These candidate cases then underwent a second screening stage using an LLM. The LLM evaluated each case against predefined criteria, including: (1) richness of clinical information (presence of history, physical examination, and laboratory findings); (2) clarity of diagnostic reasoning presented; (3) appropriateness for general medical practice settings; (4) presence of diagnostic challenges or documentation of missed diagnoses; and (5) overall educational value. This LLM-based screening prioritized common conditions presented as educational resources, atypical presentations of common conditions, and diagnostically challenging scenarios, while excluding cases focused primarily on treatment outcomes, rare genetic disorders, or highly specialized testing. We filtered for cases that match at least four out of the five criteria, and 1,272 candidate cases. From these, 1,152 vignettes were selected after an initial human review for further processing. To balance between different specialty groups, we also included cases from the MedQA dataset [24] in our dataset generation pipeline (n = 1,952).

Dataset Generation Each case that passed initial screening underwent a NLP-based transformation process to convert free-text clinical vignette into the structured sections, including patient demographics, presenting complaints, medical history, physical examination findings, laboratory results, imaging findings, and diagnostic conclusions. Each section are mapped to our schema using a LLM-based system, which was prompted with section-specific instructions and few-shot examples to identify, categorize, and map unstructured clinical information from the narrative text to the corresponding fields in our structured schema. For our screening and filtering process, we used gpt-4.1 with temperature = 0.

Clinical Validation The VivaBench dataset underwent clinical validation at multiple stages to ensure high fidelity and accuracy. Our volunteer review panel consisted of six clinicians, including one primary care physician, one emergency physician, one intensive care specialist, one anaesthetist, one ophthalmologist, and an internal medicine physician. Three medical students were also enrolled to

Journal	Specialty	Count
Cureus	General	226
J Med Case Rep	General	64
Eur Heart J Case Rep	Cardiology	31
Medicine (Baltimore)	General	25
Case Rep Med	General	18
Clin Case Rep	General	18
Case Rep Rheumatol	Rheumatology	18
J Investig Med High Impact Case Rep	General	17
SAGE Open Med Case Rep	General	17
BMC Infect Dis	Infectious Diseases	17
J Community Hosp Intern Med Perspect	General	13
Case Rep Infect Dis	Infectious Diseases	9
Case Rep Neurol	Neurology	9
J Int Med Res	Internal Medicine	8
Case Rep Emerg Med	Emergency Medicine	8
World J Clin Cases	General	8
J Med Cases	General	8
Front Neurol	Neurology	8
Case Rep Cardiol	Cardiology	7
Case Rep Endocrinol	Endocrinology	7

Table 4: Top 20 journals contributing case reports to VivaBench. The final dataset is sourced from 260 journals across all medical specialties

assist with data parsing under clinical supervision. All volunteer reviewers will be acknowledged in the final camera ready.

During the initial case screening, clinicians assessed each candidate case for clinical accuracy, appropriate difficulty level for evaluating reasoning, and educational value. After conversion to structured format, clinicians reviewed the parsed information for accuracy, completeness, and appropriate categorization of all clinical elements. Any inaccuracies, inconsistencies, and missing information were manually corrected before finalization of dataset. The list of diagnoses and accepted differentials were also manually reviewed and annotated.

C Dataset Composition

The dataset was selected from an initial pool of 250,294 PubMed case reports, involved rule-based NLP filtering (yielding $\sim 20,000$ cases), LLM-based screening (yielding 1,272 cases), and an initial human review (selecting 1,152 PubMed cases). These, along with 800 MedQA cases, were structured, creating a pool of 1,952 cases. After clinician validation, our evaluation dataset for VivaBench consists of 990 cases across nine specialty groups: Endocrine & Reproductive (150), Infectious Disease & Immunology (150), Cardiovascular & Metabolic (148), Gastrointestinal (147), Neurological/Psychiatric (136), Hematology/Oncology/Other (112), Pediatric (69), Respiratory (51), and Musculoskeletal & Pain (26). Both the complete set of 1,952 generated structured cases and the curated evaluation dataset are made available on our Hugging-Face repository: <https://huggingface.co/datasets/chychiu/VivaBench/>

The final dataset was sourced from 260 different journals, with Table 4 showing the top 20 sources by count, and their respective specialties. The journal distribution largely reflects the underlying prevalence in medical case report literature, with general medical journals predominating alongside specialty-specific publications. We note two outliers (Cureus, J Med Case Rep) from this dataset, both of which were aggregators of case reports or smaller medical publications.

Human Baseline To assess the difficulty and appropriateness of dataset in practice, we obtained a brief human baseline, where 4 clinicians provided a provisional diagnosis based on the clinical picture over 14 unseen cases. Our clinician baseline achieved a 0.68 ± 0.09 top-3 accuracy versus 0.52 ± 0.07 for models at the provisional stage, suggesting our benchmark approximates clinician-level expectations. A larger scale human baseline would be one valuable direction for future research.

model	setting	bleurt	ROUGE-F1	ROUGE-prec	ROUGE-recall
llama-4-maverick	general	0.25 +/- 0.08	0.12 +/- 0.08	0.11 +/- 0.09	0.15 +/- 0.13
	journal	0.25 +/- 0.08	0.12 +/- 0.09	0.11 +/- 0.10	0.17 +/- 0.14
grok-3-mini	general	0.25 +/- 0.07	0.12 +/- 0.09	0.13 +/- 0.10	0.14 +/- 0.11
	journal	0.25 +/- 0.06	0.12 +/- 0.08	0.13 +/- 0.09	0.12 +/- 0.08
o4-mini	general	0.22 +/- 0.08	0.10 +/- 0.09	0.11 +/- 0.10	0.12 +/- 0.11
	journal	0.21 +/- 0.07	0.10 +/- 0.07	0.10 +/- 0.09	0.12 +/- 0.09
deepseek-r1	general	0.28 +/- 0.06	0.08 +/- 0.06	0.06 +/- 0.06	0.20 +/- 0.12
	journal	0.26 +/- 0.05	0.09 +/- 0.08	0.07 +/- 0.07	0.19 +/- 0.14
qwen3-235b-a22b	general	0.24 +/- 0.07	0.10 +/- 0.07	0.09 +/- 0.08	0.13 +/- 0.10
	journal	0.24 +/- 0.06	0.10 +/- 0.08	0.10 +/- 0.09	0.13 +/- 0.09
gemini-2.5-pro	general	0.28 +/- 0.08	0.15 +/- 0.12	0.18 +/- 0.15	0.16 +/- 0.13
	journal	0.28 +/- 0.08	0.18 +/- 0.13	0.21 +/- 0.16	0.18 +/- 0.14

Table 5: The BLEURT, ROUGE-F1, precision and recall for with generic prompt (*general*) and provision of journal title (*journal*) in the prompt

C.1 Analysis on Dataset Contamination

As VivaBench was derived from existing published medical journal, there is concern dataset contamination affecting benchmark validity and results. To assess whether evaluated models had memorized complete case reports from the training data, we conducted a sentence completion experiment following the methodology of Golchin et al. [17].

Methodology We extracted the title and first sentence from each case report in our PubMed dataset and prompted models to generate the second sentence as it appeared in the original publication. We referencing the system prompt described in [17] for our system prompt:

*“You are provided with the Title and Sentence 1 from a case report from {dataset}.
Finish Sentence 2 as appeared in the dataset. Sentence 2 must exactly match the
instance in the dataset.”*

We evaluated two prompting strategies: (1) generic prompting with `dataset="PubMed database"`, and (2) journal-specific prompting using the actual source journal name. We used BLEURT [42] and ROUGE-L [32] to quantify textual similarity between generated and original sentences. In addition, we tested whether models could extract journal names from titles alone or author lists from journal and title combinations.

Results Generated sentences exhibited consistently low textual similarity to the original second sentences across all models and experimental configurations (Table 5). Manual review of the top 10% of cases ranked by either similarity metric revealed no exact matches with the original publications. Furthermore, models failed to accurately extract journal names or author lists, and frequently hallucinated references during both reasoning and output generation.

These findings suggest that while the case reports were published prior to the training data cutoff dates, the reports were unlikely to be directly memorised by the models.

D Information Retrieval and Parsing

A core technical challenge in our framework is the translation between natural language clinical queries generated by an agent and the structured information available in a clinical case $\mathcal{C}$. To address this, we developed a two-stage approach consisting of a Mapper module ($\mathbb{M}$) and a Parser module ($\mathbb{P}$).

The Mapper $\mathbb{M}$ functions as a translation mechanism. It maps any free-text clinical query $\mathcal{Q}$ originating from an agent’s action in the space $\mathcal{A} = \{\mathcal{H}, \mathcal{P}, \mathcal{I}, \mathcal{L}\}$ (representing history, physical examination, imaging, and lab investigations respectively) to a set of structured information keys $\mathcal{K}_Q \subseteq \mathcal{K}_C$, where $\mathcal{K}_C$ denotes the complete set of information keys available within a clinical case $\mathcal{C}$. For each $k_i \in \mathcal{K}_C$, there is a corresponding value $v_i \in \mathcal{V}_C$, which represents the set of all associated textual values for these keys. These key-value pairs follow the structure of our clinical schema, and defined through our data generation process.

We implemented two variants of the mapper: 1. A deterministic variant ($\mathbb{M}_D$) that maps free-text queries to relevant keys using cosine similarity embeddings, medical entity recognition, and domain-specific keyword matching. 2. An LLM-based variant ($\mathbb{M}_{LLM}$) that utilizes few-shot examples to interpret the semantic intent of queries. Our experiments use gpt-4.1 with temperature = 0 as the backbone for $\mathbb{M}_{LLM}$. Both approaches return a set of matched key-value pairs derived from $\mathcal{K}_C$ (and their associated values from $\mathcal{V}_C$) and a list of unmatched terms from the original query, thereby maintaining explicit boundaries regarding available information. Figure 7 illustrates this process with a working example.

The parser module, $\mathbb{P}$, then transforms the retrieved structured information (the matched key-value pairs from $\mathbb{M}$) into coherent natural language responses formatted according to medical documentation conventions. For history ($\mathcal{H}$) and physical examination ($\mathcal{P}$) findings, negative results (e.g., absence of a symptom) are explicitly returned if queried. Standardized laboratory values not specifically detailed in the source case $\mathcal{C}$ are returned as default normal values with appropriate reference ranges. Investigations not available in $\mathcal{C}$ are explicitly noted as "not available" to prevent information leakage or hallucination by the agent. Our framework also provides an optional LLM enhancement layer for $\mathbb{P}$ for history and physical examination outputs. This layer (using gpt-4.1) refines the natural language generation of responses, improving their fluency, while strictly adhering to the information boundaries established by the deterministic parsing logic. This ensures that responses remain factually accurate and grounded in $\mathcal{C}$, facilitating human review during evaluation.

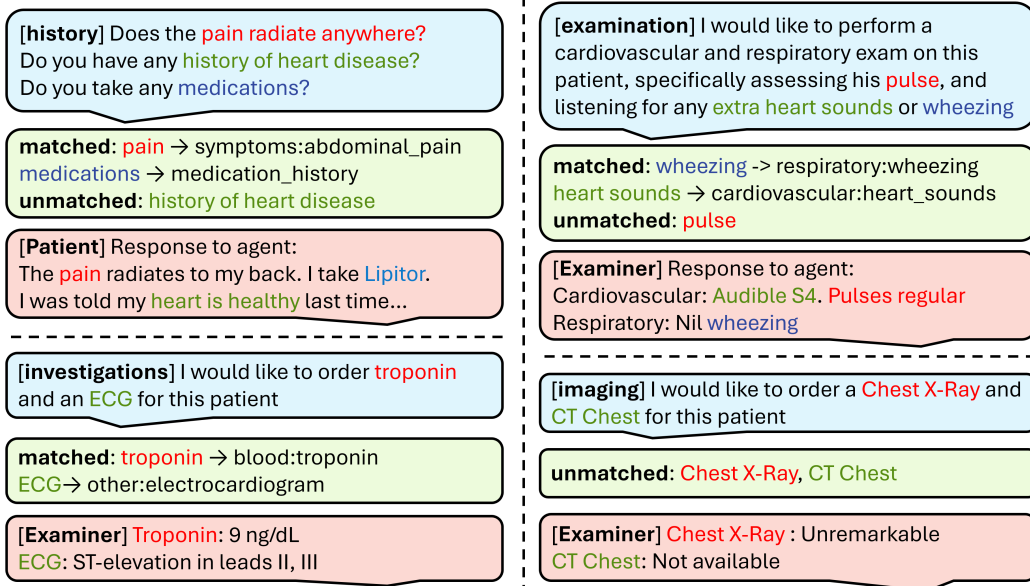


Figure 7: Example of how our Mapper module ($\mathbb{M}$, shown in green) maps specific keywords from free-text queries from agent (blue), across history ($\mathcal{H}$, top left), physical examination ($\mathcal{P}$, top right), lab investigations ($\mathcal{L}$, bottom left), and imaging ($\mathcal{I}$, bottom right). Each matched key from $\mathcal{K}_C$ has a corresponding value from $\mathcal{V}_C$, which is then processed by the Parser module ($\mathbb{P}$) into a free-text response (red). For any unmatched query terms, relevant negative findings are included for history and examination queries, and default normal values are provided for common investigations if not specified in the case (e.g., Chest X-Ray showing no acute abnormality).

D.1 Mapper Module ($\mathbb{M}$)

Deterministic Mapper ($\mathbb{M}_D$) The deterministic variant, $\mathbb{M}_D$, translates free-text queries into structured information keys through a multi-stage process detailed in Algorithm 3. For queries related to history ($\mathcal{H}$) and physical examination ($\mathcal{P}$), relevant keys from $\mathcal{K}_C$ are identified based on semantic similarity (computed using a dual-embedding approach with PubMedBERT and SNOMED-CT mediated similarity) between extracted query phrases and standardized medical terms. For laboratory ($\mathcal{L}$) and imaging ($\mathcal{I}$) requests, synonym dictionaries including terminology from medical coding systems

(e.g., LOINC), common abbreviations, and predefined investigation panels are used to recognize diagnostic tests across various terminologies.

Algorithm 3 Process for deterministic mapper $\mathbb{M}_D$

1. Medical entities $\mathcal{V}_q$ are extracted from the input free-text query Q using the SpaCy biomedical model `en_core_sci_md`.
 2. For history ($\mathcal{H}$) and physical examination ($\mathcal{P}$) queries:
 - An embedding model based on PubMedBERT embeddings (`neuml/pubmedbert-base-embeddings`) is initialized, and indexed with $\mathcal{V}_C$, the set of textual descriptions associated with keys in $\mathcal{K}_C$.
 - For each extracted query entity $v \in \mathcal{V}_q$:
 - Direct semantic similarity, $\text{sim}(v, v_c)$, is computed between v and each case-specific description $v_c \in \mathcal{V}_C$.
 - SNOMED-CT mediated similarity is computed: v is first mapped to a set of SNOMED-CT concepts S_v where $\text{sim}(v, s) > \tau_S$ (with $\tau_S = 0.8$). Then, $\text{sim}(s, v_c)$ is computed for each $s \in S_v$ and $v_c \in \mathcal{V}_C$.
 - Query entities are matched to keys in $\mathcal{K}_C$ if either their direct or SNOMED-CT mediated similarity to a v_c exceeds a predefined threshold $\tau_M = 0.6$.

This approach normalizes terminological variations (e.g., "shortness of breath" vs. "dyspnea") through medical concept alignment.
 3. For laboratory investigation ($\mathcal{L}$) and imaging ($\mathcal{I}$) queries:
 - Comprehensive synonym dictionaries, $\mathbb{D}_{LAB}$ (for labs) and $\mathbb{D}_{IMG}$ (for imaging), are utilized to map common terms, LOINC codes, laboratory abbreviations, imaging modalities, and anatomical specifications to standardized keys in $\mathcal{K}_C$.
 - For each query entity $v \in \mathcal{V}_q$, potentially relevant keys $\mathcal{K}_{relevant} = \mathbb{D}(v)$ are retrieved from the respective dictionaries.
 - Matched keys are confirmed by intersecting with keys present in the current case: $\mathcal{K}_v = \mathcal{K}_{relevant} \cap \mathcal{K}_C$.
 4. The mapper $\mathbb{M}_D$ returns:
 - A set of matched pairs: $\{(v, k) \mid v \in \mathcal{V}_q, k \in \mathcal{K}_C, \text{ and } k \text{ is matched to } v \text{ via steps 2 or 3}\}$.
 - A set of unmatched query entities: $\mathcal{U}_q = \{v \mid v \in \mathcal{V}_q, v \text{ has no match in } \mathcal{K}_C\}$.
-

LLM-based Mapper ($\mathbb{M}_{LLM}$) The LLM-based mapper, $\mathbb{M}_{LLM}$, employs an LLM (gpt-4.1 with `temperature = 0` in our experiments) as its core component to semantically parse free-text clinical queries into structured information keys. For each query Q corresponding to an action category in $\{\mathcal{H}, \mathcal{P}, \mathcal{I}, \mathcal{L}\}$, an API call is made to the LLM. The prompt includes system instructions, few-shot examples demonstrating the mapping task, the user query Q , and the set of available information keys $\mathcal{K}_C$ for the current clinical case. The LLM processes this input and returns a structured JSON response containing matched pairs $\{(v, k) \mid v \in \mathcal{V}_q, k \in \mathcal{K}_C\}$ and unmatched query entities $\mathcal{U}_q = \{v \mid v \in \mathcal{V}_q, v \text{ has no match in } \mathcal{K}_C\}$. This design allows $\mathbb{M}_{LLM}$ to handle complex or nuanced queries while preventing information leakage by grounding its output in $\mathcal{K}_C$.

To optimize $\mathbb{M}_{LLM}$, a calibration dataset of 100 clinical cases with human-annotated query-to-key mappings was used. Prompts were iteratively refined to maximize mapping accuracy against these gold-standard annotations, ensure output consistency across repeated identical requests, and improve robustness to variations in query phrasing. Examples included in system prompts were selected to cover common query patterns, including the correct handling of negative findings (e.g., "patient denies fever"), requests for normal values, and queries about unavailable information. This iterative refinement ensures that $\mathbb{M}_{LLM}$ maintains high precision in information retrieval while effectively managing the diversity of information requests encountered during agent interaction.

Validation of Mapper Modules To validate the performance and robustness of our mapper modules ($\mathbb{M}_D, \mathbb{M}_{LLM}$), we evaluated their precision (Pr) and recall (Rc) in mapping free-text queries to the correct structured information keys ($\mathcal{K}_C$). This validation was conducted on a calibration set of 100 clinical cases featuring human-annotated query-to-key mappings. Queries were sampled from those

Table 6: Performance of the Deterministic ($\mathbb{M}_D$) and LLM-based ($\mathbb{M}_{LLM}$ with gpt-4.1 backbone) Mapper variants. Precision (Pr) and Recall (Rc) were computed on a calibration set of 100 cases, using queries sampled from each of the evaluated agent models across the four information categories: $\mathcal{H}$ (History), $\mathcal{P}$ (Physical Exam), $\mathcal{I}$ (Imaging), and $\mathcal{L}$ (Labs).

Agent Model	Deterministic Mapper ($\mathbb{M}_D$)								LLM-based Mapper ($\mathbb{M}_{LLM}$)							
	$\mathcal{H}$		$\mathcal{P}$		$\mathcal{I}$		$\mathcal{L}$		$\mathcal{H}$		$\mathcal{P}$		$\mathcal{I}$		$\mathcal{L}$	
	Pr	Rc	Pr	Rc	Pr	Rc	Pr	Rc	Pr	Rc	Pr	Rc	Pr	Rc	Pr	Rc
DeepSeek-R1	0.76	0.64	0.71	0.87	0.98	0.87	0.82	0.84	0.84	0.94	0.86	0.91	1.00	0.97	0.92	0.90
o4-mini	0.78	0.68	0.75	0.90	0.94	0.80	0.79	0.87	0.89	0.94	0.88	0.96	0.98	0.94	0.89	0.85
Llama-4 Maverick	0.79	0.71	0.63	0.83	0.94	0.81	0.71	0.91	0.82	0.99	0.89	0.88	0.96	0.97	0.97	0.93
Grok 3 Mini Beta	0.77	0.67	0.67	0.90	0.96	0.84	0.77	0.91	0.88	0.96	0.92	0.97	0.98	1.00	0.97	0.94
Qwen 3 (235b-a22b)	0.82	0.64	0.75	0.95	0.94	0.78	0.75	0.88	0.84	0.94	0.87	0.95	0.98	0.97	0.95	0.88
Gemini 2.5 Pro	0.77	0.59	0.72	0.81	0.97	0.78	0.80	0.87	0.83	0.94	0.90	0.92	1.00	0.96	0.94	0.83

generated by each of the evaluated agent models across the four clinical categories ($\mathcal{H}, \mathcal{P}, \mathcal{I}, \mathcal{L}$). Specifically, for each agent model, 100 queries were sampled for each category to serve as input to $\mathbb{M}_D$ and $\mathbb{M}_{LLM}$.

Table 6 outlines the result our validation. Overall, both mapper variants are largely consistent when processing queries generated by different agent models. This suggests that the mappers do not exhibit significant bias towards the query style of any particular agent architecture, or for $\mathbb{M}_{LLM}$, its base LLM model family (i.e. o4-mini vs. others). The $\mathbb{M}_{LLM}$ variant generally showed higher Pr/Rc, particularly for $\mathcal{H}$ and $\mathcal{P}$ queries. This was likely due to limitations on entities extraction process in $\mathbb{M}_D$, where it is unable to handle the richness and complexity in descriptive free-text queries.

Furthermore, to confirm the operational determinism of $\mathbb{M}_{LLM}$ for identical inputs, we conducted an additional test where 100 distinct queries were each submitted 10 times to $\mathbb{M}_{LLM}$ (with a fixed case context). This experiment yielded an Intersection over Union (IoU) of >0.99 for the set of mapped keys and a variance of <0.01 in the mapped output across all repetitions. These findings confirm that $\mathbb{M}_{LLM}$, despite using a LLM to assist with information processing, produces highly deterministic and reliable mappings for identical queries within a given clinical case context. For the rest of our experiments, we use $\mathbb{M}_{LLM}$ due to its superior performance over $\mathbb{M}_D$.

To assess inter-rater reliability, we conducted a post-hoc validation using two independent clinicians on 50 cases, with a weighted Cohen’s kappa of 0.655 (moderate agreement).

D.2 Parser Module ($\mathbb{P}$)

After the Mapper module ($\mathbb{M}_D$ or $\mathbb{M}_{LLM}$) translates an agent’s query into a set of structured (key, value) pairs corresponding to information available in $\mathcal{K}_C$, a Parser module ($\mathbb{P}$) transforms these retrieved clinical values into coherent natural language responses. For each action category ($\mathcal{H}, \mathcal{P}, \mathcal{I}, \mathcal{L}$), $\mathbb{P}$ formats the matched clinical information using pre-set templates designed to emulate standard medical documentation. $\mathbb{P}$ also maintains a record of previously processed keys within an interaction session to prevent redundant information retrieval and presentation.

For history ($\mathcal{H}$) and physical examination ($\mathcal{P}$) findings, unmatched requests are considered negative results, and are returned as relevant negatives if the information is not available in $\mathcal{C}$ (e.g., "Query: Does the patient have a fever? Response: Negative for fever."). For common investigations in $\mathcal{I}, \mathcal{L}$ not specifically detailed in $\mathcal{C}$, default normal values are returned within predefined reference ranges. Uncommon lab investigations or imaging requests for data not available in $\mathcal{C}$ are explicitly reported as "not available" to prevent agent hallucination.

The framework also includes an optional LLM-based enhancement layer that can be applied to the output of $\mathbb{P}$ specifically for history and physical examination responses. This layer utilizes the same LLM as $\mathbb{M}_{LLM}$ (gpt-4.1). For history queries, it simulates patient responses in a first-person perspective, maintaining appropriate medical literacy. For physical examination queries, it formats findings as concise clinical notes organized by body system. This LLM enhancement improves the naturalness and readability of responses for human review, while strictly adhering to the information boundaries and factual content established by the core deterministic parsing logic.

E Evaluation Details

Evaluated Models We evaluated several state-of-the-art foundation models, including Gemini 2.5 pro [18], DeepSeek-R1 [13], o4-mini from OpenAI [36], Llama-4 Maverick [33], Grok 3 mini beta [48], and Qwen 3 (235b-a22b) [8]. These models were selected based on their performance ranking on the Artificial Analysis Leaderboard [9]. Each model was accessed through OpenRouter [4] and tested at `temperature = 0`.

Evaluation Task Each structured clinical case $\mathcal{C}$ in our VivaBench benchmark consists of five components: history ($\mathcal{H}$), physical examination ($\mathcal{P}$), imaging investigations ($\mathcal{I}$), laboratory investigations ($\mathcal{L}$), and a set of ground truth diagnoses ($\mathcal{D}$) with accepted differentials ($\mathcal{D}'$).

Models interacted with these cases through a free-flowing dialogue system. They could request information from any category using specific actions (e.g., `history`, `examination`, `investigation`, `imaging`) and received responses synthesized from the structured case content. Models were instructed to first gather patient information through history-taking and physical examination, then provide a provisional diagnosis (d_P) before ordering any diagnostic tests. Subsequently, they were to order necessary investigations and, upon determining they had sufficient information, provide a final diagnosis (d_D). The complete system prompt detailing these instructions and available actions is provided in Algorithm 4.

We imposed a global turn limit of 20 interaction steps per case. Category-specific request limits were also enforced: 10 for history ($\mathcal{H}$), 5 for physical examination ($\mathcal{P}$), 3 for imaging ($\mathcal{I}$), and 3 for laboratory ($\mathcal{L}$) requests. Each information request counted as one step towards the global limit. If an agent reached a category-specific limit, it was instructed to proceed to the next phase of the diagnostic process. Failed requests due to formatting errors could be retried up to twice per original request; these retries counted towards the global action limit but not the category-specific limits.

Our evaluation assessed model performance at two key stages:

- **Provisional Stage:** After the initial review of patient history and physical examination findings, but before ordering any investigations, the model provided one or more provisional diagnoses (d_P) with associated confidence values.
- **Final Stage:** After ordering and reviewing relevant diagnostic investigations, the model provided one or more final diagnoses (d_D) with updated confidence values.

For both provisional and final diagnoses, models were instructed to provide up to five possible diagnoses in a specified JSON format, each including: a free-text condition name, the corresponding ICD-10 name, the ICD-10 code, and a confidence score (0.0-1.0).

Additionally, we conducted a **Full Information** pre-test. In this setting, models were provided with the complete clinical picture (all information from $\mathcal{H}, \mathcal{P}, \mathcal{I}, \mathcal{L}$) upfront, bypassing the interactive information-gathering process. This served as a control to differentiate knowledge recall and synthesis capabilities from active information-seeking and sequential reasoning abilities.

Evaluation Data and Metrics The complete output trace for each model interaction—including all queries, retrieved information (matched keys), provisional diagnoses, and final diagnoses—was preserved. All evaluated models achieved a first-pass success rate (i.e., completed the interaction without critical errors) of over 97% on the VivaBench dataset. For subsequent metric computation, we used the intersection of successfully completed cases across all models, resulting in an evaluation set of 934 cases. The specific metrics computed for diagnostic accuracy, confidence calibration, information seeking efficiency, and diagnostic adaptation are detailed in Appendix F. Full experimentation logs are available upon request.

Computational Cost The token usage and approximate costs associated with executing one full run of the VivaBench benchmark (evaluating one model over 934 cases) are summarized in Table 7. On average, evaluating one model requires 26.4M tokens on average, with estimated costs ranging from \$11.9 to \$389 depending on the specific LLM’s pricing at the time of experimentation. These cost figures can be considered the upper bound for each model, as we did not consider the price difference in input / output tokens, or account for potential cost reductions from caching mechanisms for repeated prompts or API calls in our calculations.

Algorithm 4 System prompt for AI agent for evaluation task (Examples omitted)

You are a primary care medical AI assistant. You are currently reviewing a patient. Your goal is to perform a full diagnostic workup for the patient, and find the underlying diagnosis to the patient's presentation.

Workflow constraints:

1. You must first gather patient information through history and examination before ordering any tests
2. After reviewing the patient at bedside, you should provide a provisional diagnosis, before ordering any investigations
3. Once you order any lab or imaging investigations, you can no longer gather additional history or perform examinations on the patient
4. You can only perform one action at a time
5. When you have sufficient information, you should provide a final diagnosis

Available actions:

- **history:** Interview the patient directly. Ask only 1-2 questions at a time to avoid overwhelming them. Assume average medical literacy.
- **examination:** Perform a physical examination. Specify exactly what examination you want to perform and what signs you're looking for.
- **diagnosis_provisional:** Provide your provisional diagnosis given a clinical picture, after reviewing the patient but before ordering any investigations or imaging.
- **investigation:** Order any investigations that are not imaging. If you are ordering a laboratory test, specify which laboratory tests you are ordering, and specimen type if the laboratory test you are ordering is not serological. Bedside tests such as ECG, and other special tests, such as EEG, Pulmonary Function Tests etc., go here as well.
- **imaging:** Order medical imaging. Imaging modalities are strictly limited to imaging modalities that are performed by a radiologist, radiographer, or nuclear medicine physician, such as xray, ultrasound, CT, MRI, PET-scan etc. VQ scan also included here. Specify both the modality and anatomical region.
- **diagnosis_final:** Provide your final diagnosis after completing your evaluation.

For diagnoses (both provisional and final):

- Some patients might have multiple issues/diagnoses, or you may not be certain about this patient's diagnosis. You may list up to five possible diagnoses if there are multiple or if you are uncertain.
 - For each diagnosis, provide the condition name, ICD-10 name, ICD-10 code, and your confidence (0.0-1.0) about the diagnosis. The condition name can be any descriptive text you choose, while the ICD-10 name needs to adhere to ICD-10 terminology.
 - Confidence scores do not need to sum to 1.0
 - **Format as a list of dictionaries:** `[{"condition": "free text name of the condition", "icd_10_name": "icd 10 name of the condition", "icd_10": "icd code of the condition", "confidence": score}]`
 - Remember to always give your provisional diagnosis before ordering any investigations
-

Table 7: Approximate token usage and costs per model for a full VivaBench evaluation run (934 cases). "Per Case Tokens (k)" indicates the mean number of tokens (in thousands) used per case, including standard deviation. Costs are estimated based on OpenRouter pricing prevailing during experiments.

Model	Success (%)	Total Tokens (M)	Per Case Tokens (k)	Rate (\$/M)	Est. Cost (\$)
Gemini 2.5 Pro	97.1	38.9	40.4 ± 8.3	10.00	389.00
DeepSeek-R1	98.6	20.0	20.5 ± 3.8	2.50	50.00
o4-mini	98.1	31.0	31.9 ± 9.2	4.40	136.40
Llama-4 Maverick	99.9	22.7	23.0 ± 4.6	0.60	13.60
Grok 3 Mini Beta	100.0	23.8	24.1 ± 3.5	0.50	11.90
Qwen 3 (235b-a22b)	100.0	21.9	22.1 ± 4.0	0.60	13.10

Table 8: Top-3 accuracy of DeepSeek-R1 on VivaBench with varying temperature over exact (-E) and approximate (-A) diagnoses in provisional (d_P), final diagnosis (d_D), and full-information (d_F) stages

t	d_P -E	d_D -E	d_F -E	d_P -A	d_D -A	d_F -A
0	0.27 +/- 0.02	0.46 +/- 0.03	0.81 +/- 0.01	0.53 +/- 0.03	0.71 +/- 0.03	0.98 +/- 0.01
0.1	0.28 +/- 0.04	0.46 +/- 0.04	0.80 +/- 0.02	0.51 +/- 0.02	0.73 +/- 0.03	0.99 +/- 0.01
0.3	0.27 +/- 0.04	0.47 +/- 0.03	0.81 +/- 0.01	0.53 +/- 0.03	0.73 +/- 0.01	0.98 +/- 0.01
0.5	0.26 +/- 0.03	0.45 +/- 0.03	0.81 +/- 0.01	0.52 +/- 0.01	0.70 +/- 0.03	0.97 +/- 0.01
1	0.27 +/- 0.02	0.42 +/- 0.04	0.81 +/- 0.02	0.52 +/- 0.03	0.69 +/- 0.03	0.98 +/- 0.01

E.1 Temperature settings and statistical reliability

For systematic evaluation of model performance over clinical reasoning, we chose temperature 0 to improve reproducibility of our results. While there might be tradeoffs (e.g. temp 0 being greedy decoding), this allows us to isolate the models’ deterministic reasoning capabilities from stochasticity introduced in higher temperature ranges. To further explore impact of temperature on benchmark results, we repeated experiments with temperature variations during benchmark.

Methodology: We re-evaluated four models (DeepSeek-R1, grok-3-mini, Llama-4 Maverick, Qwen-235) across temperature settings of [0, 0.1, 0.3, 0.5, 1] with 5 independent runs each, on a subset of 100 cases from our evaluation set. This subset was selected to preserve the accuracy distribution compared to our benchmark across the models being evaluated. We measured top-3 approximate and exact accuracies under the same experimental setup as our main experiments.

Results: Across all models, mean diagnostic accuracy remained stable across temperature settings, with differences typically within 2-3 percentage points from the temperature 0 baseline ($p > 0.05$ for all comparisons). There is some increase in variance and decrease in accuracy at higher temperatures, suggesting decreased consistency with increasing temperature. Illustrative example for DeepSeek-R1 outlined in Table 8 above.

Upon qualitative analysis, we observed that models tend to succeed or fail on the same cases consistently across runs, even at higher temperatures. Review of agent trajectories showed that while the specific reasoning paths varied at higher temperatures, the diagnostic outcomes remained largely unchanged for individual cases. This suggests that the models’ clinical reasoning capabilities are the primary determinant of performance on our benchmark. These findings align with recent work by Renze et al [38], who found that temperature variations in reasoning tasks often introduce surface-level changes in explanation style without substantially altering underlying logical capabilities.

F Evaluation Metrics

F.1 Diagnostic Accuracy

We assessed diagnostic accuracy using several complementary metrics, focusing on the models’ ability to identify correct diagnoses within their top-k predictions at different stages of the evaluation (Provisional d_P , Final d_D , and Full Information pre-test).

Table 9: Expanded Top-k **Approximate** Diagnostic Accuracy across different evaluation stages: Provisional diagnosis (d_P), Final diagnosis (d_D after interaction), and Full Information pre-test (d_F). These approximate accuracy metrics complement the exact accuracy results presented in Table 2 in the main paper.

Model	Provisional (d_P)			Final (d_D)			Full Information (d_F)		
	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
Gemini 2.5 Pro	0.65	0.70	0.72	0.75	0.79	0.80	0.93	0.98	0.98
DeepSeek-R1	0.61	0.66	0.67	0.71	0.75	0.75	0.93	0.97	0.97
o4-mini	0.62	0.67	0.68	0.74	0.77	0.77	0.94	0.98	0.98
Llama-4 Maverick	0.59	0.64	0.64	0.70	0.71	0.71	0.91	0.96	0.96
Grok 3 Mini Beta	0.61	0.66	0.68	0.65	0.70	0.72	0.92	0.96	0.96
Qwen 3 (235b-a22b)	0.56	0.60	0.61	0.68	0.72	0.72	0.90	0.96	0.96

Top-k Exact Accuracy We determined whether any of an agent’s top-k predictions exactly matched a ground truth diagnosis in $\mathcal{D}$. A prediction was considered an exact match if its specified condition name or ICD-10 name/code precisely matched an entry in $\mathcal{D}$. For the Provisional (d_P) and Final (d_D) interactive stages, an additional criterion for an exact match was that the agent must have also ordered at least one relevant investigation supporting that specific diagnosis; this criterion does not apply to the Full Information stage. Exact accuracy results are presented in Table 2 (main paper).

Top-k Approximate Accuracy We also assessed whether any of an agent’s top-k predictions approximately matched a ground truth diagnosis in $\mathcal{D}$ or an accepted differential diagnosis in $\mathcal{D}'$. A prediction was considered an approximate match if it met any of the following conditions:

- It matched the ICD-10 code of any diagnosis in $\mathcal{D}$ at a broader hierarchical level (e.g., predicting I23 for a ground truth of I23.1).
- It matched the ICD-10 name or condition name of any accepted differential diagnosis listed in $\mathcal{D}'$.
- Its ICD-10 name or condition name exhibited high semantic similarity (cosine similarity > 0.8) to any diagnosis in $\mathcal{D}$ or $\mathcal{D}'$.

For the Provisional (d_P) and Final (d_D) diagnostic stages, a prediction qualifying as an approximate match under these criteria also required the agent to have ordered at least one relevant investigation supporting that diagnosis or differential. This investigation ordering criterion does not apply to the Full Information (d_F) stage. Expanded Top-k Approximate Accuracy results are presented in Table 9.

Multiple Diagnosis Handling For clinical cases in $\mathcal{C}$ with multiple ground truth diagnoses specified in $\mathcal{D}$, accuracy (both exact and approximate) was counted as a success (value of 1.0) if any one of the ground truth diagnoses was correctly identified by the agent within its top-k predictions, according to the respective criteria.

F.2 Confidence Metrics

For each diagnosis item provided by an agent in its provisional (d_P) and final diagnosis (d_D) lists, a confidence score (0.0 to 1.0) was required.

Raw Confidence Values We tracked the mean raw confidence scores across all diagnoses provided by an agent at the provisional stage (C_P in Table 2) and final diagnosis stage (C_D in Table 2).

Confidence-Weighted Accuracy Score (S_{conf}) To evaluate how well an agent’s confidence aligns with its diagnostic correctness, we first normalized the confidence scores (c_j) for all N diagnoses proposed by the agent in a given stage for a case, such that $\sum_{j=1}^N c'_j = 1.0$, where c'_j is the normalized confidence. We then calculated S_{conf} as:

$$S_{\text{conf}} = \sum_{i \in E} c'_i + \sum_{i \in A} c'_i - \sum_{i \in U} c'_i$$

Where c'_i is the normalized confidence for the i -th prediction, and E, A, U are the sets of exact, approximate, and unmatched diagnostic predictions respectively. This score ranges from -1.0 to +1.0.

Results for the provisional and final diagnostic stages are presented as S_P and S_D in Table 2 of main paper respectively.

Confidence-Accuracy Matrix To explore the relationship between diagnostic performance and confidence calibration, we mapped models’ confidence against accuracy using kernel density estimation (KDE) distributions (Figure 5, main paper). For each clinical case, a model’s mean raw confidence score (y-axis) was plotted against its composite diagnostic accuracy (x-axis; defined as average of Top-k exact and approximate accuracies across $k \in \{1, \dots, 5\}$). KDE contours (bandwidth 0.7, 50% density mass) were calculated to visualize case distribution per model. The matrix quadrants (Confidently Accurate, Overconfidently Wrong, Underconfident but Accurate, Appropriately Uncertain) denote distinct patterns in how different models calibrate uncertainty against their knowledge within our simulated clinical setting.

F.3 Information Seeking Efficiency

We evaluated how models strategically gather clinical information by measuring *Information seeking efficiency*, which we define as the ability to selectively request relevant history ($\mathcal{H}$), physical examination ($\mathcal{P}$), and diagnostic investigations ($\mathcal{I}, \mathcal{L}$) that contribute to an accurate diagnosis. To measure this, we tracked binarised metrics of matched information keys in our evaluation:

$$\text{Precision} = \frac{|K_{\text{requested}} \cap K_{\text{relevant}}|}{|K_{\text{requested}}|} \quad \text{Recall} = \frac{|K_{\text{requested}} \cap K_{\text{relevant}}|}{|K_{\text{relevant}}|} \quad (1)$$

Where $K_{\text{requested}}$ is the set of clinical information elements requested by the model, and K_{relevant} is the set expert-annotated as relevant for diagnosis. Metrics were averaged across cases. Calculations used four configurations (Table 10), based on information stage (Review vs. Investigation) and diagnostic scope (Matched vs. Full). In general, "Matched Scope" assesses targeted inquiry; "Full Scope" evaluates thoroughness. Precision reflects inquiry focus, while Recall indicates completeness.

Table 10: Configurations for Information Seeking Efficiency Metrics.

	Matched Scope (Targeted to proposed diagnoses)	Full Scope (All diagnosis-relevant information)
Review Stage	Precision and recall of history ($\mathcal{H}$) and physical examination ($\mathcal{P}$) items relevant <i>only</i> to the diagnoses actively proposed by the agent.	Precision and recall of <i>all</i> potentially relevant history ($\mathcal{H}$) and physical examination ($\mathcal{P}$) items for the case, regardless of the agent’s current diagnostic hypotheses.
Investigation Stage	Precision and recall of imaging ($\mathcal{I}$) and laboratory ($\mathcal{L}$) investigations relevant <i>only</i> to the diagnoses actively proposed by the agent.	Precision and recall of <i>all</i> potentially relevant imaging ($\mathcal{I}$) and laboratory ($\mathcal{L}$) investigations for the case, regardless of the agent’s current diagnostic hypotheses.

F.4 Diagnostic Evolution Metrics

To further scrutinize the reasoning process within the evaluated models, we analyzed how their diagnostic hypotheses and associated confidence levels evolved between the provisional (d_p) and final (d_f) stages of the interaction. This examination aims to determine whether an agent’s capacity to adapt its diagnostic judgment ("change its mind") and recalibrate its confidence in response to new information (from investigations) correlates with material improvements in diagnostic performance.

Outcome Metrics To quantify the net change in diagnostic performance from the provisional to the final stage, we measured the following outcome metrics:

- **Exact Accuracy Change:** The change in Top-1 Exact Accuracy (Final d_f - Provisional d_p). (Corresponds to "Acc Chg." in Table 3).
- **Approximate Accuracy Change:** The change in Top-1 Approximate Accuracy (Final d_f - Provisional d_p). (Corresponds to "Apr. Acc Chg." in Table 3).
- **S_{conf} Change:** The change in the Confidence-Weighted Accuracy Score (S_{conf}) (Final d_f - Provisional d_p). (Corresponds to " S_{conf} Chg." in Table 3).

Variable Metrics To characterize the specific adjustments made by models to their diagnostic hypotheses, we tracked the following variable metrics:

- **Diagnoses Added/Removed/Maintained:** The counts of unique diagnoses that were added to, removed from, or maintained in the agent’s list of top-5 hypotheses when transitioning from the provisional to the final stage. (Corresponds to "Diag. Added", "Diag. Removed", "Diag. Maintained" in Table 3).
- **Confidence Delta:** The overall change in mean raw confidence across all proposed diagnoses, calculated as (Mean Raw Confidence at Final Stage, C_D) - (Mean Raw Confidence at Provisional Stage, C_P). (Corresponds to "Conf. Delta ($C_D - C_P$)" in Table 3).
- **Confidence Shift:** For diagnoses that were maintained in the agent’s list from the provisional to the final stage, this is the mean of the difference in their confidence scores ($c_{f,j} - c_{p,j}$ for each maintained diagnosis j). (Corresponds to "Conf. Shift" in Table 3).
- **Confidence Shift Magnitude:** For diagnoses maintained between stages, this is the mean of the absolute difference in their confidence scores ($|c_{f,j} - c_{p,j}|$ for each maintained diagnosis j). (Corresponds to "Conf. Shift Mag." in Table 3).

Correlation Analysis We analyzed the statistical relationships between these diagnostic adaptation variable metrics and the observed outcome metrics. The objective was to identify whether specific patterns of adjusting diagnostic hypotheses (e.g., adding/removing diagnoses, magnitude of confidence shifts) systematically correlate with improvements in final diagnostic accuracy or confidence calibration. For categorical outcome variables (e.g., whether exact accuracy improved), Welch’s t-test (for two groups) or one-way ANOVA (for more than two groups) was applied to the variable metric distributions. For continuous outcome variables (e.g., the numerical change in S_{conf}), Pearson correlation coefficients were computed between the outcome and variable metrics. The results of this correlation analysis are presented and discussed in the main paper (Table 3). Broadly, these findings suggest that more effective diagnostic performance is often associated with a greater willingness to adapt initial hypotheses and recalibrate confidence in light of new evidence from diagnostic investigations.

G Qualitative Analysis

Below, we include several examination traces, including agent reasoning, action, and query, as examples of potential failure modes we observed in our evaluation. Any mentioned relevant diagnoses are highlighted in purple, and specific sections that reflect the issues are highlighted in red.

Premature Diagnostic Closure In a case involving a 28-year-old male presenting with epigastric pain, the agent correctly identified two potential diagnoses from clinical review, pancreatitis (confidence score 0.7) and peptic ulcer disease (confidence score 0.2), and ordered the appropriate tests for pancreatitis. Laboratory results revealed elevated amylase and lipase, which are positive for pancreatitis. This prompted the agent to conclude its journey, and diagnosed the pancreatitis without further investigations. However, the agent failed to pursue testing for the underlying etiology. In this case, it was due to a stricture from a duodenal ulcer, which could have been identified through a H. pylori testing, which is a routine test for any patients with suspected peptic ulcer[15]. This reflects satisfaction-of-search bias, where identification of one diagnosis prematurely terminates the diagnostic process, potentially missing causative or concomitant conditions.

Examiner: Clinical case stem: 28 year old male presenting with epigastric pain.
Vitals: Heart Rate: 92. Blood Pressure: 122/78. Temperature: 37.1. Respiratory rate: 16. Oxygen saturation: 99.

Please review and diagnose the patient.

Agent: The patient is a young adult male presenting with epigastric pain and stable vital signs. To narrow down the differential diagnosis, I need more information about the characteristics of the pain, such as onset, duration, quality, and any aggravating or relieving factors.

[history] Can you describe the epigastric pain? When did it start, and is it constant or does it come and go?

Examiner: The pain started suddenly earlier today. It's in the upper middle part of my abdomen, and it's been severe and constant since it began. It doesn't really come and go - it's just always there. Nothing seems to make it better or worse.

Agent: Sudden, severe, constant epigastric pain in a young male could be due to several causes, including peptic ulcer disease (with or without perforation), acute pancreatitis, biliary pathology, or even less common causes like myocardial ischemia or referred pain. I need to clarify associated symptoms to narrow the differential.

[history] Have you experienced any nausea, vomiting, fever, or changes in your bowel movements? Have you noticed any blood in your vomit or stool?

Examiner: I have been feeling nauseous, but I haven't had any vomiting or fever. My bowel movements haven't changed, and I haven't noticed any blood in my vomit or stool.

Agent: I need to further characterize the pain and assess for risk factors or associated symptoms that could help narrow the differential. I want to know if the pain radiates anywhere and if the patient has any history of similar pain, alcohol use, or known medical conditions.

[history] Does the pain radiate to your back, chest, or shoulders? Have you ever had similar pain before? Do you drink alcohol or have any known medical conditions such as ulcers or gallstones?

Examiner: The pain does radiate to my back. I've never had similar pain before. I drink alcohol occasionally - about two cans of beer twice a week. I don't have any known medical conditions, and I've never had ulcers or gallstones as far as I can tell.

Agent: The sudden onset of severe, constant epigastric pain radiating to the back raises concern for acute pancreatitis, but other causes such as peptic ulcer disease (including perforation), biliary pathology, or even less likely cardiac causes should be considered. I need to examine the abdomen for tenderness, guarding, rebound, and any palpable masses to further narrow the differential.

[examination] I want to perform an abdominal examination, specifically checking for tenderness, guarding, rebound tenderness, palpable masses, and any signs of peritonism.

Examiner: Abdomen:

Tenderness: Epigastric tenderness with mild guarding.

Guarding: Mild guarding present.

Rebound tenderness: No rebound tenderness.

Palpable masses: No palpable masses.

Peritonism: No signs of peritonism.

Agent: The patient is a young male with sudden-onset, severe, constant epigastric pain radiating to the back, associated with nausea but no vomiting, fever, or GI bleeding. Examination reveals epigastric tenderness with mild guarding but no peritonism or masses. The most likely causes include **acute pancreatitis, peptic ulcer disease (possibly with early perforation or severe gastritis)**, and less likely biliary or cardiac causes given the location and radiation. The absence of peritonism and stable vitals make perforation less likely, but acute pancreatitis is a strong consideration given the pain's character and radiation.

[diagnosis provisional] (**condition: Acute pancreatitis, confidence: 0.7**), (**condition: Peptic ulcer disease, confidence: 0.2**), (condition: Gastritis, confidence: 0.1)

Examiner: Thank you. Please proceed to imaging and lab investigations.

Agent: Acute pancreatitis is the leading provisional diagnosis given the sudden severe epigastric pain radiating to the back, nausea, and epigastric tenderness with mild guarding. To confirm this and assess severity, I will order serum amylase and lipase, as well as liver function tests to rule out biliary involvement.

Additionally, an abdominal ultrasound is indicated to assess for gallstones or pancreatic abnormalities.

[investigation] Order serum amylase, serum lipase, and liver function tests (ALT, AST, ALP, GGT, bilirubin).

Examiner: Blood:

Amylase: 1145.0 U/L

Lipase: 1103.0 U/L

Alanine Aminotransferase (ALT): 45.0 IU/L

Aspartate Aminotransferase (AST): 43.0 IU/L

Alkaline Phosphatase (ALP): 97.0 IU/L

Gamma-Glutamyl Transferase (GGT): 149.0 U/L

Total Bilirubin: 0.64 mg/dL

Agent: The patient's clinical presentation of sudden severe epigastric pain radiating to the back, along with significantly elevated serum amylase and lipase, is highly consistent with **acute pancreatitis**. Liver function tests are only mildly elevated, and there is no evidence of biliary obstruction or hepatitis. Other differentials are much less likely given the lab findings and clinical picture. **[diagnosis final]** (condition: Acute pancreatitis, confidence: 1.0)
Diagnosis: Acute pancreatitis secondary to duodenal ulcer-induced ampullary obstruction, **acquired duodenal stenosis from chronic duodenal ulcer**

Inadequate investigations In a case of cerebral infarct, the agent was presented with a 63-year-old male with facial droop, numbness, and significantly elevated blood pressure (212/123 mmHg). The agent correctly suspected a pontine infarct in its provisional diagnosis from the clinical signs alone, but only ordered a non-contrast CT scan, which had low sensitivity (approximately 41%) of CT imaging for posterior fossa lesions such as pontine strokes[23]. The imaging only showed chronic ischemic changes, and the agent diagnosed the patient with a transient ischemic attack (TIA). While a non-contrast CT scan was first-line for stroke patients, this case illustrates how models follow procedural pathways without appropriately weighing test limitations against clinical presentation.

Stem: 63/M presenting with facial droop and other non-specific neurological symptoms . HR 98, BP 212/123 mmHg. Please review and diagnose the patient.

Agent: The patient presents with facial droop and neurological symptoms alongside significantly elevated blood pressure, suggesting a potential stroke or hypertensive emergency. I need to gather more information about the patient's symptoms and history to narrow down the differential diagnosis.

[history] Can you describe the facial droop and other neurological symptoms you're experiencing? When did they start?

Patient: It started suddenly about three hours ago. At first, it was on the right side of my face, but that went away by the time I got here. Then, while I was waiting, I felt some weakness on the left side across my whole face. I also felt a bit numb on the left side of my body, but it shifted to the right face and arm afterwards.

Agent: The migratory nature of the patient's symptoms is concerning for a transient ischemic attack (TIA) or evolving stroke. The significantly elevated blood pressure could be either a cause or consequence of cerebrovascular pathology. I should enquire other associated symptoms.

[history] Have you experienced any headache, vision changes, dizziness, confusion, or trouble with balance or coordination?

Patient: I also have some difficulty swallowing, and my wife said my speech was a bit slurred. However, I don't feel it now.

Agent: The patient's symptoms of sudden onset facial droop, shifting numbness, difficulty swallowing, and slurred speech suggest a neurological condition. The fluctuating nature of the symptoms could indicate a complex or dynamic pathology. I need to examine the patient to assess neurological deficits and look for signs that could help narrow down the diagnosis.

[examination] I want to perform a neurological examination to assess cranial nerve function, muscle strength, sensation, and reflexes, looking for signs of stroke, demyelinating disease, or other neurological conditions.

Examiner: Neurological: Cranial nerves grossly intact, no facial asymmetry currently visible. Muscle strength normal in all limbs (5/5 bilaterally). Sensation decreased to pinprick and cold temperature on left side, particularly in the left upper and lower extremities, consistent with the patient's reported symptoms. Reflexes 2+ and symmetrical throughout. Gait appears normal with no ataxia. Coordination intact on finger-to-nose and heel-to-shin testing. No pronator drift observed. NIH Stroke Scale score 1 for mild left-sided sensory deficit.

Agent: This patient presents with acute onset left facial paralysis (upper and lower face), left facial sensory loss, dysarthria, left-sided cerebellar signs, and a severe headache, with a history of transient left then right-sided numbness. There is no limb weakness or sensory loss in the limbs. The blood pressure is severely elevated. The findings localize to the left lateral pons or medulla, suggesting a brainstem syndrome, most likely a lateral pontine (AICA) or lateral medullary (Wallenberg) syndrome. The severe hypertension raises concern for hypertensive emergency with possible brainstem ischemia or hemorrhage. Other differentials include demyelinating disease or a mass lesion, but the acute presentation and vascular risk factors make stroke most likely."

[provisional diagnosis] (condition: Acute ischemic stroke, left lateral pontine or medullary syndrome, confidence: 0.7), (condition: Hypertensive emergency with possible brainstem involvement, confidence: 0.4), (condition: Intracerebral hemorrhage, confidence: 0.2), (condition: Multiple sclerosis, confidence: 0.1)

Examiner: Thank you. Please proceed to imaging and lab investigations.

Agent: The patient's presentation is highly suggestive of an acute brainstem event (ischemic or hemorrhagic stroke). An urgent non-contrast CT head is needed to rule out hemorrhage, and MRI brain with diffusion-weighted imaging is ideal for detecting acute ischemia. Given the hypertensive emergency, basic labs are also warranted to assess end-organ damage.

[imaging] Order a **non-contrast CT scan** of the head to assess for acute hemorrhage or large infarct.

Examiner: == CT BRAIN ==

Chronic periventricular ischemic changes without evidence of acute ischemia or hemorrhage

Agent: **Imaging does not show acute infarct or hemorrhage**, but the patient has focal neurological deficits (left facial paralysis, left facial sensory loss, left-sided cerebellar signs, dysarthria) and hypertensive emergency. The findings are most consistent with an acute brainstem (pontine) syndrome, possibly a **transient ischemic attack (TIA) or a small vessel (lacunar) event not visible on imaging**.

Demyelinating disease is less likely given the age and risk factors, but cannot be fully excluded. The chronic ischemic changes support a vascular etiology.'

[investigation] I would like to check this patient's lipid profile, HbA1c, and a thrombophilia screen to assess for risk factors for stroke and potential underlying causes.

Examiner: Investigation Results

Total cholesterol: 245 mg/dL

LDL cholesterol: 168 mg/dL

HDL cholesterol: 38 mg/dL

Triglycerides: 195 mg/dL

HbA1c: 6.8%

Thrombophilia screen: Pending

Agent: The patient has focal neurological deficits (left facial paralysis, left facial sensory loss, dysarthria, left-sided cerebellar signs) with a history of fluctuating symptoms and severe hypertension. Imaging shows only chronic ischemic changes, with no acute infarct or hemorrhage. The clinical picture is most consistent with a brainstem (pontine) transient ischemic attack (TIA) or a small vessel (lacunar) syndrome, likely related to hypertensive small vessel disease. There is no evidence of acute stroke or hemorrhage on imaging, and labs are unremarkable. The most likely diagnosis is a hypertensive brainstem TIA, but a small vessel lacunar syndrome remains possible. Demyelinating disease is less likely given the age and lack of other features.

[final diagnosis] (condition: **Transient ischemic attack (TIA), brainstem, confidence: 0.7**), (condition: Hypertensive encephalopathy (with small vessel disease), confidence: 0.5), (condition: Lacunar syndrome (chronic small vessel ischemic disease), confidence: 0.3)

Diagnosis: Acute medial pontomedullary junction infarction

Inappropriate Hypothesis Generation When presented with a 4-month-old infant with feeding difficulties, one agent inappropriately prioritized heart failure in its differential diagnosis, stating: "To differentiate potential heart failure from respiratory pathology, I need more detail on feeding tolerance, sweating, cyanosis, and progression of breathing difficulty." This led to subsequent ordering of inappropriate tests that are routine for heart failure, while failing to investigate other possible causes of the patient's presentation. This demonstrates a fundamental misalignment between the agent's reasoning and patient-specific clinical probabilities, suggesting limitations in the model's ability to appropriately contextualize presentations within pediatric epidemiology.

Stem: 4 week old female presenting with poor feeding and trouble breathing. Temperature 37.5 C, HR 172 bpm, BP 90/50 mmHg, RR 32/min, O2 sat 100% Please review and diagnose the patient.

Agent: To differentiate potential heart failure from respiratory pathology, I need more detail on feeding tolerance, sweating, cyanosis, and progression of breathing difficulty.

[history] How many feeds does she take per day and how much volume each time? Does she sweat or turn blue when feeding or crying?

Examiner: She's been feeding a lot less than usual - maybe just a few times a day, and she only takes a small amount each time, not finishing her usual bottles. She doesn't sweat or turn blue when she's feeding or crying.

Agent: We need to characterize the breathing difficulty - whether it's tachypnea, retractions, grunting, apneic spells, or upper airway issues - to narrow the differential.

[history] Can you describe her trouble breathing? Does she have rapid breathing, grunting, nasal flaring, chest retractions, coughing, or any pauses in breathing (apnea)?

Examiner: She seems to be having trouble breathing, but I haven't noticed her breathing fast, grunting, flaring her nostrils, pulling in at her chest, coughing, or having any pauses where she stops breathing.

Agent: Poor feeding, tachycardia, and reported breathing difficulty raise concern for cardiac pathology or **early heart failure**. I need to check for murmurs, gallops, and hepatomegaly.

[examination] I want to listen to this patient's heart sounds for any murmurs or gallop rhythms and palpate the abdomen for hepatomegaly.

Examiner: Cardiovascular: Heart sounds regular, no murmur or gallop rhythm. Abdominal: No hepatomegaly.

Agent: The caregiver reports breathing difficulty despite no overt distress; I need to assess the lungs for any subtle abnormalities.

[examination] I want to auscultate the patient's lungs for any crackles, wheezing, or diminished breath sounds, and inspect the chest for symmetry and any deformity.

Examiner: Respiratory: Bilateral crackles on auscultation. No wheezing. Breath sounds not diminished. Chest inspection normal, no asymmetry or deformity.

Agent: The infant's poor feeding, tachycardia, and bilateral crackles without primary respiratory distress signs are most consistent with **early congestive heart failure** from a congenital cardiac lesion, with pneumonia less likely.

[diagnosis provisional] (condition: **Congestive heart failure**, likely secondary to congenital heart defect, confidence: 0.7), (condition: Ventricular septal defect, confidence: 0.2), (condition: Pneumonia, unspecified organism, confidence: 0.1)

Examiner: Thank you. Please proceed to imaging and lab investigations.

Agent: A chest X-ray will help assess heart size, pulmonary vascular markings, and evidence of pulmonary edema or structural abnormalities suggestive of a congenital heart lesion.

[imaging] I would like to order a Chest X-Ray (PA and lateral)

Examiner: == X-RAY CHEST ==

Chest distension without focal infection

Agent: To evaluate for heart failure severity, rule out infection, and assess for possible myocarditis, I will order blood tests including markers of inflammation and cardiac stress plus cultures and a viral panel.

[investigation] Complete Blood Count; C-Reactive Protein; **B-type Natriuretic Peptide (serum)**; Blood cultures; Viral respiratory PCR panel (nasopharyngeal swab)

Examiner: Blood:

- White Blood Cell Count: 9.2 -> 18.4 x 10⁹/L - on admission, after 48 hours
- Hemoglobin: 10.9 -> 10.8 g/dL - on admission, after 48 hours
- Platelet Count: 553 -> 734 x 10⁹/L - on admission, after 48 hours
- C-Reactive Protein: <5 mg/L - on admission and after 48 hours
- Mean corpuscular volume: Normal
- B-type natriuretic peptide: Normal

Microbiology:

- Blood Culture, Aerobic: negative
- Blood Culture, Anaerobic: negative
- Respiratory Viral Panel: negative - including RSV and influenza

Agent: The presentation of poor feeding, respiratory crackles, chest distension on X-ray without infection, and signs of pulmonary overcirculation raise suspicion for an underlying structural cardiac abnormality **despite absence of a murmur**; echocardiography is the next best step to evaluate for congenital heart disease or cardiomyopathy.

[imaging] I would like to order a transthoracic echocardiogram

Examiner: Echocardiogram not available.

Agent: An electrocardiogram can reveal signs of cardiomyopathy or structural heart disease (e.g., chamber enlargement, conduction abnormalities) when echocardiography is unavailable.

[investigation] Electrocardiogram (ECG)

Examiner: Other:

- Electrocardiogram: Normal

Agent: The infant's presentation of poor feeding, tachycardia, bilateral pulmonary crackles on exam, chest distension on X-ray, and absence of infection or primary pulmonary disease is **most consistent with congestive heart failure** from an underlying congenital cardiac lesion. Without echocardiography, the precise defect is uncertain but a large ventricular septal defect or total anomalous pulmonary venous return remain possible.

[diagnosis final] (condition: Congestive heart failure, confidence: 0.9), (condition : Ventricular septal defect, confidence: 0.4), (condition: Total anomalous pulmonary venous connection, confidence: 0.2)

NeurIPS Paper Checklist

1. Claims

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

Answer: [\[Yes\]](#)

Justification: The abstract and introduction outlines the motivation and general direction of our paper

Guidelines:

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

2. Limitations

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

Answer: [\[Yes\]](#)

Justification: Limitations was discussed in the paper.

Guidelines:

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

3. Theory assumptions and proofs

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

Answer: [\[NA\]](#)

Justification: Our results are mostly experimental

Guidelines:

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

4. Experimental result reproducibility

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

Answer: [\[Yes\]](#)

Justification: We provide the dataset and code necessary to replicate our experiments in our submission.

Guidelines:

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

5. Open access to data and code

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

Answer: [Yes]

Justification: Dataset and code are accessible. Croissant uploaded

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

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

Answer: [Yes]

Justification: Brief experimental settings are outlined within the paper. The full experiment setting is described in the appendix.

Guidelines:

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

7. Experiment statistical significance

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

Answer: [Yes]

Justification: Notations signifying statistical significance is included in the paper

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

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

Answer: [Yes]

Justification: We provide the full technical details in our appendix

Guidelines:

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

9. Code of ethics

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

Answer: [Yes]

Justification:

Guidelines:

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

10. Broader impacts

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

Answer: [Yes]

Justification: This has potential impact on the implementation of clinical LLMs

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

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

Answer: [NA]

Justification: Our dataset is publicly sourced with patient information deidentified

Guidelines:

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

12. Licenses for existing assets

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

Answer: [Yes]

Justification: The licenses of datasets we used are all open domain / creative-commons

Guidelines:

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

- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. **New assets**

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

Answer: [\[Yes\]](#)

Justification:

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

Answer: [\[NA\]](#)

Justification:

Guidelines:

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

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

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

Answer: [\[NA\]](#)

Justification:

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [Yes]

Justification: [Yes]

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.