Note2Chat: Improving LLMs for Multi-Turn Clinical History Taking Using Medical Notes

Yang Zhou[1*†], Zhenting Sheng[2*], Mingrui Tan[1], Yuting Song[1], Jun Zhou[1], Yu Heng Kwan[3,4], Lian Leng Low[3,4], Yang Bai[1†], Yong Liu[1]

[1]Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore
[2]Nanyang Technological University
[3]National University of Singapore
[4]Singapore General Hospital

Abstract

Effective clinical history taking is a foundational yet underexplored component of clinical reasoning. While large language models (LLMs) have shown promise on static benchmarks, they often fall short in dynamic, multi-turn diagnostic settings that require iterative questioning and hypothesis refinement. To address this gap, we propose Note2Chat, a note-driven framework that trains LLMs to conduct structured history taking and diagnosis by learning from widely available medical notes. Instead of relying on scarce and sensitive dialogue data, we convert real-world medical notes into high-quality doctor-patient dialogues using a decision tree-guided generation and refinement pipeline. We then propose a three-stage fine-tuning strategy combining supervised learning, simulated data augmentation, and preference learning. Furthermore, we propose a novel single-turn reasoning paradigm that reframes history taking as a sequence of single-turn reasoning problems. This design enhances interpretability and enables local supervision, dynamic adaptation, and greater sample efficiency. Experimental results show that our method substantially improves clinical reasoning, achieving gains of +16.9 F1 and +21.0 Top-1 diagnostic accuracy over GPT-4o.

Code — https://github.com/zhentingsheng/Note2Chat

Introduction

History taking and differential diagnosis are fundamental to clinical reasoning, forming the basis for understanding a patient's condition and directing subsequent diagnostic and therapeutic decisions. A thorough history typically encompasses the chief complaint, history of present illness, review of systems, and general medical and social background, all of which collectively inform the generation and refinement of a differential diagnosis, a ranked list of plausible conditions grounded in the patient's symptoms and risk factors. Central to this process is *multi-turn clinical history taking*, a dynamic, interactive dialogue in which clinicians iteratively ask targeted questions, interpret responses in context, and update diagnostic hypotheses step by step (Henderson, Tierney, and Smetana 2012; Guyatt et al. 2015). This sequential reasoning demands broad medical knowledge, adaptability, and time—resources often constrained in high-volume care environments. Despite these challenges, history taking alone can lead to accurate diagnoses in a majority of cases (Kuriakose 2020; Nierenberg 2020), underscoring its diagnostic value. In this context, automated history-taking systems hold significant promise: by conducting structured, multi-turn interviews prior to clinical encounters, they can streamline information gathering, reduce cognitive load on clinicians, and enhance the completeness and consistency of diagnostic conversations, particularly in settings with limited medical personnel.

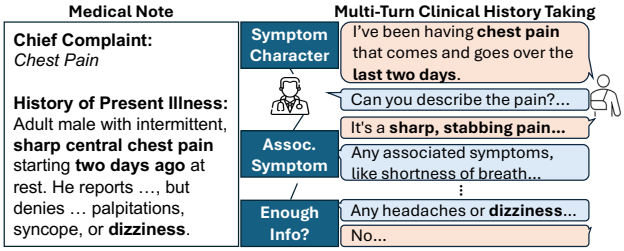

Figure 1: Multi-turn clinical history taking is the dynamic process of gathering information from a patient; structured medical notes are the organized product, synthesizing that narrative into a concise, standardized record.

Large language models (LLMs) have demonstrated impressive performance across a range of medical tasks, including medical question answering, clinical note summarization, and care plan generation (Cabral et al. 2024; Goh et al. 2024; McDuff et al. 2025; Nori et al. 2023a,b, 2024). However, these achievements are largely based on static, single-turn benchmarks, where models are provided with complete clinical vignettes and tasked with producing an answer without needing to interact or inquire further. Such settings fail to capture the sequential and exploratory nature of real-world diagnostic reasoning, which requires actively gathering missing information through dialogue. Recent benchmark studies have underscored this limitation, show-

*These authors contributed equally.
†Corresponding Authors.

ing that when LLMs are evaluated in full diagnostic conversations, where they must initiate questions, adapt based on responses, and iteratively refine hypotheses, their diagnostic accuracy can drop significantly compared to single-turn tasks (Johri et al. 2025; Liu et al. 2024; Li et al. 2024; Hager et al. 2024; Schmidgall et al. 2024). This performance gap indicates that, despite their medical knowledge, current LLMs lack the conversational competencies needed for effective multi-turn diagnostic reasoning. In particular, they often fail to generate focused follow-up questions or to prioritize clinically relevant details, constraining their usefulness in structured interviews (Goh et al. 2024; Nori et al. 2024). These findings point to the need for dynamic, interaction-oriented evaluations that better reflect the challenges of history taking and conversational diagnosis.

Recent efforts to enhance LLMs for clinical dialogue have explored self-play simulation, agent-based workflows, and reinforcement learning (RL) strategies. For example, AMIE (Tu et al. 2025) introduced a simulated diagnostic environment to improve history-taking dialogue, but relies on private datasets and models. DoctorAgent-RL (Feng et al. 2025) has introduced RL fine-tuning to encourage LLMs to progressively refine diagnoses through proactive questioning. Agent-based methods (Nori et al. 2025; Gatto et al. 2025; Liu et al. 2025; Rose et al. 2025) assign different roles to separate LLMs, yet typically use general-purpose models not tailored for clinical reasoning. RL-based approaches (Fansi Tchango et al. 2022; Sun et al. 2025) aim to improve diagnostic performance through fine-tuning, but often depend on rigid supervision or task-specific annotations, limiting adaptability. Despite their differences, most of these works prioritize final diagnosis accuracy and underemphasize the quality and completeness of history taking. They often overlook clinically important but non-diagnostic details, such as negative findings or symptom context, and are limited by the lack of large-scale, high-quality dialogue data. Given the unique challenges of history taking, including its exploratory nature and variation across clinical styles, there is a clear need for scalable and generalizable training paradigms.

In this work, we address the challenges of clinical history taking by focusing on efficient information gathering rather than optimizing solely for diagnostic accuracy. Our objective is to enable the model to extract as many relevant findings as possible with minimal questioning, supporting concise and complete interviews grounded in clinical reasoning. To this end, we propose `Note2Chat`, a novel framework that leverages real-world medical notes, specifically the primary diagnosis and history of present illness (HPI), as a supervision signal. These notes capture clinician-curated summaries of symptom relevance, temporal progression, and diagnostic thinking, offering a rich and widely accessible resource for training. Compared to medical dialogues, clinical notes are significantly more available, as they are routinely documented for care delivery and are less restricted by privacy concerns. Moreover, they require no additional manual annotation and can be easily adapted to local protocols and institutional practices.

Our framework consists of three core components designed to enhance LLMs for clinical history taking. First, we introduce a **note-to-dialogue generation pipeline** that converts discharge notes into clinically meaningful doctor-patient conversations using decision tree-guided prompts, followed by refinement to ensure realism and comprehensive coverage of key findings. Using this pipeline, we construct a dataset comprising 8,944 synthetic dialogues, 67,077 successful rollouts, and 11,403 preference pairs across 4,972 patients. Second, we propose a **three-stage fine-tuning strategy** that combines supervised training on note-guided dialogues, data augmentation through simulated interactions, and direct preference optimization (DPO) (Rafailov et al. 2023) to encourage concise and clinically effective conversations. Third, we introduce a **single-turn reasoning paradigm** that treats each dialogue turn as an independent decision step, enabling the model to make context-aware, interpretable actions guided by conversation history and reasoning plans. This design improves follow-up questioning, streamlines information gathering, and allows for explicit reward shaping based on information gain and diagnostic relevance. Our contributions include:

- We propose a note-driven dialogue generation and refinement pipeline to curate clinically grounded patient-doctor interactions, resulting in a history-taking dataset across 4,972 patients.
- We propose a three-stage fine-tuning strategy and a single-turn reasoning paradigm that enhance LLMs for concise, interpretable, and effective history taking.
- Using `Note2Chat`, our fine-tuned LLM achieves state-of-the-art performance in both information gathering and diagnosis, with relative gains of +57.53% and +42.86% over GPT-4o, respectively.

Related Work

LLMs for medical applications: LLMs have demonstrated impressive capabilities across a broad range of medical applications, including question answering, clinical summarization, and care planning (Cabral et al. 2024; Goh et al. 2024; McDuff et al. 2025; Nori et al. 2023a,b, 2024; Achiam et al. 2023; Saab et al. 2024). Advanced reasoning models such as DeepSeek-R1 (Guo et al. 2025) and Gemini 2.5 (Comanici et al. 2025) continue to raise the bar with stronger generalization and reasoning performance on medical benchmarks. Meanwhile, domain-specialized variants like BioMistral (Labrak et al. 2024), HuatuoGPT-o1 (Chen et al. 2025a), Med-R1 (Lai et al. 2025), and MedGemma (Sellergren et al. 2025) further tailor LLMs to clinical contexts by incorporating structured medical knowledge and fine-tuning on healthcare-specific data. Despite these advances, most evaluations remain confined to static, single-turn settings in which models are presented with complete case information and asked to produce a response. Such paradigms overlook the inherently dynamic and sequential nature of real-world clinical reasoning—particularly in tasks like history taking and differential diagnosis, where success hinges on actively collecting missing information, ask-

ing follow-up questions, and reasoning under uncertainty. Recent studies have shown that even state-of-the-art models struggle in these interactive, multi-turn environments, with diagnostic performance dropping significantly when required to reason step-by-step without full context (Johri et al. 2025; Liu et al. 2024; Li et al. 2024; Hager et al. 2024; Schmidgall et al. 2024). These limitations underscore the need for new training and evaluation frameworks that support proactive, reasoning-driven dialogue and reflect the complexities of real clinical workflows.

LLMs for multi-turn clinical conversation: Recent research has increasingly focused on enabling LLMs to participate in multi-turn clinical conversations, particularly for tasks such as history taking and differential diagnosis. Several benchmarks (Johri et al. 2025; Li et al. 2024; Schmidgall et al. 2024; Wang et al. 2025; Chandra et al. 2025; Fan et al. 2024) have been introduced to evaluate LLMs in interactive medical settings, offering useful tools for assessing question-asking and reasoning abilities. Multi-agent frameworks that assign roles like history taker or diagnostician to separate LLMs aims to improve the workflow rather than LLMs' clinical reasoning ability. RL-based methods (Fansi Tchango et al. 2022; Sun et al. 2025) promote proactive questioning and diagnostic refinement, but depend on rigid, pre-defined state-action spaces that limit flexibility and generalizability. AMIE (Tu et al. 2025) marks a notable advancement by training models in a self-play diagnostic environment and extending to multimodal data (Saab et al. 2025), yet it relies on proprietary datasets and closed-source models, limiting reproducibility and broader adoption. To address data scarcity, prior curation efforts (Tu et al. 2025; Fansi Tchango et al. 2022; Saley et al. 2024; Chen et al. 2025b) generate training data from annotated dialogues, synthetic vignettes, or QA-style conversions, and typically use supervised fine-tuning (SFT) for training. DoctorAgent-RL (Feng et al. 2025) further applies RL fine-tuning to enhance question generation but remains constrained by synthetic, limited training resources.

Note2Chat

We introduce **Note2Chat**, a generalizable LLM training framework for clinically grounded, proactive history taking. By leveraging medical notes as a natural and scalable supervision source, our approach shifts the focus from diagnosis to high-quality information gathering, better aligning with the exploratory nature of real-world clinical reasoning.

Problem Setup. We formulate medical history taking as a *partially observable sequential decision-making process*, where a doctor agent interacts with a simulated patient grounded in clinical notes. The objective of the doctor agent is to elicit clinically relevant findings through follow-up questions and ultimately produce a differential diagnosis.

Let $\mathcal{P}$ denote the distribution over patient cases. Each case $x \sim \mathcal{P}$ is defined as $x = \{\text{dx}, \mathcal{F}, \text{cc}\}$, where dx is the ground-truth diagnosis extracted from the note, $\mathcal{F} = \{f_1, \ldots, f_n\}$ is the set of clinical findings extracted from the *History of Present Illness (HPI)*, and cc is the chief complaint, serving as the initial observation. At each turn t, the doctor agent observes a state $s_t = \{\text{cc}, h_t\}$, where $h_t = [(q_1, r_1), (q_2, r_2), \ldots, (q_{t-1}, r_{t-1})]$ is the accumulated dialogue history, where q_t and r_t are the doctor question and patient response, respectively. The agent selects an action $a_t \in \mathcal{A} = \mathcal{A}^{\text{ask}} \cup \mathcal{A}^{\text{diagnose}}$, either asking a follow-up question q_t or issuing a diagnostic prediction.

If the agent selects a question-asking action $a_t \in \mathcal{A}^{\text{ask}}$, a simulated patient grounded in x responds with $r_t \sim p(r \mid q_t, x)$, and the dialogue history is then updated to include the new exchange, resulting in the next state $s_{t+1} = \{\text{cc}, h_{t+1}\}$, where $h_{t+1} = h_t \cup \{(q_t, a_t)\}$. Alternatively, if $a_t \in \mathcal{A}^{\text{diagnose}}$, the interaction terminates and the doctor agent will predicta ranked list of K potential diagnoses $\hat{\mathbf{y}}_t = [\hat{y}_t^{(1)}, \hat{y}_t^{(2)}, \ldots, \hat{y}_t^{(K)}] \subset \mathcal{Y}$, where $\mathcal{Y}$ is the space of diagnostic labels. Each interaction forms a trajectory $\tau = (s_0, a_0, s_1, a_1, \ldots, s_T, a_T)$, ending when a diagnosis is made. The doctor agent learns a policy $\pi_\theta(a_t \mid s_t)$, parameterized by θ, governs decision-making. The learning objective is to train the policy π_θ to maximize the expected reward $R(h_T)$, which evaluates the informativeness and efficiency of the dialogue: $\max_\theta \mathbb{E}_{x \sim \mathcal{P}, \pi_\theta}[R(\tau)]$. This can be optimized via supervised fine-tuning (SFT), reinforcement fine-tuning, or preference-based fine-tuning guided by note-derived supervision.

Data Curation Pipeline. We develop a data curation pipeline to generate clinically grounded doctor-patient dialogues from medical notes, consisting of three key components. **Finding extraction:** We extract relevant medical findings from the HPI section of discharge notes to construct patient vignettes, excluding downstream information such as lab results, treatments, and follow-up plans that would not be known during history taking. These vignettes serve as the basis for generating patient responses and for evaluating whether a model's question can recover the findings documented in the original note. **Decision tree-guided dialogue generation:** To ensure clinical relevance and alignment with diagnostic reasoning, we construct a decision tree that maps findings to candidate diagnoses. This tree provides a structured outline for guiding the LLM to generate task-oriented dialogues that reflect realistic differential diagnosis workflows. **Critic and revision:** While LLM-generated dialogues are generally plausible, they may omit key findings or exhibit context leakage, where the doctor infers symptoms not yet revealed by the patient. To improve quality, we introduce an LLM-based critic that identifies and corrects these issues by adding missing questions and revising premature inferences, significantly improving dialogue quality and increasing average symptom coverage.

To prepare the data for training and evaluation, we use ICD-10 codes to select discharge notes from the MIMIC-IV dataset (Johnson et al. 2023), focusing on two major condition groups: heart failure, cellulitis, and their associated diseases. This results in a diverse set of 10 clinically relevant conditions: Asthma, COPD, Cellulitis, Chronic venous insufficiency, Deep vein thrombosis, Erysipelas, Heart Failure, Necrotising Fasciitis, Pneumonia, Trauma/hematoma.

These conditions were selected based on clinical guidance to balance diagnostic challenge and feasibility, as they often present with overlapping symptoms that demand careful history taking to differentiate. To ensure data quality, we include only notes with a clearly defined HPI section, a primary diagnosis, and at least 100 words to guarantee sufficient clinical detail. Using GPT-4o for data processing, our curation pipeline produces multi-turn dialogues averaging 17.8 turns for 4,972 patients across 10 diseases, which are randomly split into 4,472 for training and 500 for testing.

Three-Stage Fine-Tuning Strategy. We propose a three-stage fine-tuning strategy to train LLMs for high-quality multi-turn history taking and differential diagnosis. **Cold start with SFT:** We initialize training using Qwen2.5-7B (Team 2024) as the base model. Using our note-guided dialogues, we apply supervised fine-tuning to teach the LLM foundational clinical reasoning and dialogue structure. The fine-tuned model plays the doctor role and interacts with a simulated patient agent (Qwen2.5-32B), learning to follow an appropriate question-asking flow and identify relevant findings. This stage establishes a basic starting policy for structured clinical interactions.

Self-augmentation with trajectory sampling: Note-guided dialogues are often overly idealized, with each doctor question reliably eliciting a relevant response, which is unrealistic in practice. As a result, models trained solely on these dialogues may overfit and struggle to generalize during inference. To improve robustness, we simulate more diverse and imperfect interactions by allowing the SFT-trained doctor model to engage in self-play with the patient agent. For each case, we roll out multiple dialogue trajectories and select those that achieve correct differential diagnoses with the highest recall (i.e., capturing the most documented findings). These selected dialogues are added to the training corpus, exposing the model to more natural conversation dynamics. This process yields 4,472 self-augmented dialogues from 67,077 successful rollouts.

Direct preference optimization: While supervised fine-tuning offers a solid starting point, it does not explicitly teach the model to prefer concise, effective, and clinically sound conversations. To address this, we apply DPO to guide the model toward preferred history-taking behaviors. For each case, we generate 15 dialogue candidates per case via self-play and assign a reward score to each, based on information recall, diagnostic accuracy, and dialogue efficiency. Preference pairs are then constructed by contrasting top- and bottom-ranked dialogues, and the model is optimized to favor high-quality interactions. This step strengthens the model's ability to ask relevant questions, avoid unnecessary turns, and make timely, well-justified diagnoses.

Dialogue-level outcome reward: Designing an effective reward function is critical to the success of DPO. We introduce a *dialogue-level* reward function that leverages the medical note as a silver-standard reference, explicitly capturing three desirable criteria: (1) thorough information gathering, (2) concise and efficient dialogue, and (3) accurate differential diagnosis. Formally, the reward for a dialogue trajectory τ is defined as:

$$R(\tau) = \text{Recall} + \frac{\text{Recall}}{\text{Recall}_{\text{max}}} \cdot \left(1 - \frac{\text{rank}(\text{dx}, \hat{\mathbf{y}}_T)}{K}\right) - \frac{\alpha \cdot T}{2}. \tag{1}$$

Here, Recall $= \frac{|\mathcal{F}_\tau \cap \mathcal{F}|}{|\mathcal{F}|}$, measures the fraction of clinically relevant findings $\mathcal{F}$ (from the HPI) that are successfully elicited during the dialogue τ, directly rewarding thorough information gathering. $\text{Recall}_{\text{max}}$ denotes the highest recall achieved across all generated dialogues. The second term assesses diagnostic accuracy by checking whether the ground-truth diagnosis dx appears within the top-K predicted diagnoses $\hat{\mathbf{y}}_T$; higher rankings yield higher scores. In this work, we set $K = 5$ and assign $R(\tau) = 0$ when the correct diagnosis falls outside the top-K. Importantly, this term is weighted by the ratio $\frac{\text{Recall}}{\text{Recall}_{\text{max}}}$, ensuring that high diagnostic accuracy only contributes meaningfully to the reward if sufficient relevant information has been collected, preventing reward inflation from lucky guesses. The final term $\frac{\alpha \cdot T}{2}$ penalizes unnecessarily long dialogues to encourage efficiency in history taking, where T represents the total number of turns, and α is a scaling coefficient that balances the penalty relative to the other reward components.

To generate preference pairs for DPO, we roll out 15 dialogue candidates per patient through self-play and compute the mean (μ) and standard deviation (σ) of their reward scores. Dialogues with scores above $\mu + \sigma$ are labeled as high-quality, while those below $\mu - \sigma$ are considered low-quality. For each high-quality trajectory, we sample up to two low-quality ones to form training pairs. This results in 11,403 dialogue-level preference pairs used to fine-tune the model toward more informative, accurate, and efficient multi-turn clinical conversations.

Multi-Turn History Taking via Single-Turn Reasoning. While multi-turn DPO aims to improve conversational models, it presents inherent limitations. A primary challenge is the difficulty of controlling long conversational rollouts, where early errors can compound and lead the dialogue off-track. The preference signal in conventional multi-turn approaches is applied to the entire trajectory, offering only coarse supervision that makes it difficult to identify and correct specific suboptimal turns. Furthermore, without a clear reward signal for a process like reaching a diagnosis, multi-turn models may struggle with poorly defined stopping criteria, leading to inefficient and unnecessarily long dialogues.

Task reformulation: To address these challenges, we propose a novel paradigm that reframes multi-turn history taking as *a sequence of single-turn reasoning problems*. Our core insight is that a multi-turn conversation can be modeled as a Markov Decision Process. By enriching the state representation at each turn to include the accumulated memory of prior turns and collected findings, we enable the model to make optimal, context-aware decisions from any intermediate step. This approach allows for stronger local supervision and improved sample efficiency, as the model's questioning strategy can dynamically adapt throughout the dialogue. At each step, the model should perform structured reasoning

through a thought process that includes **memory summarization, clinical planning, and action generation**. This ensures that each decision is based on the current state alone, decoupling the model's performance from the unpredictable nature of long-term conversational trajectories.

Single-turn reasoning: To achieve this, we insert a structured reasoning block between each round of patient–doctor dialogue, generated by the doctor agent during inference but hidden from the patient agent. This block, delimited by `<think>` and `</think>` tokens, comprises two key components: Summary: A concise, structured summary of the conversation history, which is updated with each new patient response. This serves as the model's working memory of the patient's condition and collected symptoms. Planning: A clear clinical rationale for the next action. If the action is a follow-up question, the plan explains how it helps to confirm or rule out specific diagnoses. If the action is a diagnostic decision, the plan justifies the ranking of candidate diseases based on the symptoms gathered. This structured approach enables more granular and flexible preference learning, where individual turns can be rewarded even if the full trajectory is suboptimal. The explicit reasoning process also enhances both interpretability and reasoning quality, as each decision is grounded in a transparent clinical rationale. A concrete example for the proposed single-turn reasoning is shown below:

Single-Turn Reasoning Example

Doctor
`<think>`
Summary: Turn 0: The patient reported having a fever and shortness of breath.
Plan: I need to understand when the symptoms started and how they've progressed to decide what might be causing them.
`</think>`
Can you tell me when your symptoms started and how they've changed?
Patient (Turn 1)
The symptoms started a few days ago and have been getting worse.
Doctor
`<think>`
Summary: Turn 0: The patient reported fever and shortness of breath.
Turn 1: The patient said the symptoms started a few days ago and have worsened.
Plan: I need to ask more about the shortness of breath to understand its nature and possible causes.
`</think>`
Can you tell me more about the shortness of breath? Has anything helped or made it worse?

Turn-level process reward: The single-turn reasoning paradigm allows for a simple, interpretable reward function at each step of interaction:

$$R_{\text{ST}}(s_{t-1}, s_t) = \begin{cases} \mathbb{I}[f_t \in s_t \setminus s_{t-1}], & \text{if } a_t \in \mathcal{A}^{\text{ask}} \\ \text{Recall}_t \cdot \left(1 - \frac{\text{rank}_t}{K}\right), & \text{if } a_t \in \mathcal{A}^{\text{diagnose}} \end{cases} \tag{2}$$

Here, $\mathbb{I}[f_t \in s_t \setminus s_{t-1}]$ indicates whether a *new relevant finding* was acquired at turn t; $\text{Recall}_t = \frac{|\mathcal{F}_t \cap \mathcal{F}|}{|\mathcal{F}|}$ measures the proportion of ground-truth findings $\mathcal{F}$ recovered so far; and $\text{rank}_t = \text{rank}(\text{dx}, \hat{\mathbf{y}}_t)$ gives the rank of the correct diagnosis in the top-K predictions. We assign $R_{\text{ST}}(s_{t-1}, s_t) = 0$ when the correct diagnosis falls outside the top-K. The agent is rewarded for acquiring new information and ranking the correct diagnosis higher, with rewards scaled by the amount of useful information gathered. Unlike multi-turn approaches, this setup makes question-asking and diagnosis actions directly comparable, enabling the model to learn when to stop asking and make a prediction. The single-turn reasoning paradigm decomposes complex dialogues into independent, context-aware decisions, allowing for precise supervision, interpretable reasoning, and verifiable rewards. This structure supports flexible preference learning and more accurate diagnoses, even when full dialogues are noisy or imperfect.

Single-turn data preparation: Starting from our self-augmented dialogue dataset, we decompose full dialogues into individual turns and use Qwen2.5-32B to generate a structured reasoning block for each of them. Following the multi-turn DPO setup, we adopt a multi-stage training strategy: first fine-tuning an LLM on the single-turn augmented data, then using the model to roll out 10 candidate responses per turn by interacting with a simulated patient agent. After filtering out low-quality samples, we retain 80,537 context-aware single-turn interactions. To construct preference data for single-turn DPO, we contrast the highest- and lowest-reward responses (as defined in Eq. 2), yielding 95,811 turn-level preference pairs.

Experiments

We conduct a series of experiments to evaluate the performance of LLMs in medical history taking and differential diagnosis. Our evaluation begins by comparing `Note2Chat`, trained with either multi-turn (`Note2Chat-MT`) or single-turn (`Note2Chat-ST`) DPO strategies, against a diverse set of baseline LLMs. These include: *Proprietary models:* GPT-4o (Achiam et al. 2023), o4-mini, Gemini-2.5 (Comanici et al. 2025); *Public open-source models:* DeepSeek-R1 (Guo et al. 2025), Qwen2.5 (Team 2024), Qwen3 (Team 2025); and *Domain-specific models:* HuatuoGPT-o1(Chen et al. 2025a), MedGemma (Sellergren et al. 2025), DoctorAgent-RL (Feng et al. 2025). We then analyze the key factors that contribute to effective history taking across clinically meaningful symptom categories. Finally, we validate our models by comparing with practicing clinicians on a held-out test set. Our experiments are designed to answer the following research questions: (1) *How well do existing LLMs perform in medical history taking?* (2) *Can we improve LLMs to proactively ask follow-up questions and autonomously decide when to diagnose?* (3) *To what extent can fine-tuning and preference learning narrow the performance gap?* (4) *How does the performance of* `Note2Chat` *compare to that of human clinicians?*

Experiments Setup. Following CRAFT-MD (Johri et al. 2025) and its prompting strategy, we simulate a patient

Model	F1	Recall	Precision	Top-1	Top-2	Top-3	#Turn
GPT-4o	29.2	33.2	30.5	49.0	61.4	67.6	22.9
o4-mini	23.0	28.7	21.9	47.6	60.0	67.0	27.0
Gemini-2.5-flash	26.6	35.5	26.7	51.4	66.2	73.0	31.9
Qwen2.5-7B-Instruct	19.6	15.7	33.0	38.8	54.8	63.2	10.3
Qwen3-8B	17.9	13.8	34.1	33.4	46.6	55.2	8.9
DeepSeek-R1-0528-Qwen3-8B	29.6	34.0	32.7	37.2	51.6	61.2	23.4
HuatuoGPT-o1-8B	0.2	0.1	1.1	19.4	33.0	42.8	2.02
MedGemma-4B-it	27.2	31.6	28.0	40.6	55.2	62.2	23.4
MedGemma-27B-text-it	27.9	31.4	30.1	52.8	66.2	71.4	21.4
DoctorAgent-RL	28.4	35.1	27.5	35.6	-	-	26.4
`Note2Chat-MT`	<u>43.8</u>	**55.4**	<u>41.8</u>	<u>62.0</u>	<u>78.2</u>	<u>82.6</u>	27.5
`Note2Chat-ST`	**46.1**	<u>46.2</u>	**54.5**	**70.0**	**81.2**	**84.4**	17.3

Table 1: History taking and diagnosis performance (%) across different models (**Best**, <u>Second Best</u>)

Mode	Model	F1	Recall	Precision	Top-1	Top-2	Top-3	Avg.	Δ	#Turn
	Qwen2.5-7B-Instruct	19.6	15.7	33.0	38.8	54.8	63.2	37.5	-	10.3
MT	+SFT	32.6	30.1	44.9	53.0	66.0	73.0	49.9	+12.4	14.1
	+SFT+Self-Aug	40.6	39.2	<u>50.1</u>	<u>62.8</u>	75.0	81.6	58.2	+20.7	15.8
	+SFT+Self-Aug+DPO	<u>43.8</u>	**55.4**	41.8	62.0	<u>78.2</u>	<u>82.6</u>	<u>60.6</u>	<u>+23.1</u>	27.5
ST	+SFT	35.4	37.5	40.7	54.8	65.0	70.4	50.6	+13.1	19.8
	+SFT+Self-Aug	41.4	44.8	45.8	60.8	72.4	75.4	56.8	+19.3	20.5
	+SFT+Self-Aug+DPO	**46.1**	<u>46.2</u>	**54.5**	**70.0**	**81.2**	**84.4**	**63.7**	**+26.2**	17.3

Table 2: Ablation study showing the impact of each component in our `Note2Chat` framework (**Best**, <u>Second Best</u>).

agent using Qwen2.5-32B, which interacts with the evaluated LLMs acting as doctor agents. The doctor agent is tasked with asking relevant follow-up questions to elicit key clinical findings and terminating the conversation once sufficient information has been gathered. After the interview, the doctor outputs a ranked list of potential diagnoses for differential diagnosis. To evaluate performance, we use a Qwen2.5-32B model to assess the dialogue. It checks how many ground-truth findings from the note are successfully elicited in the conversation and computes the rank of the true diagnosis within the predicted list. We report the precision, recall, F1 scores, and Top-K accuracy as our primary evaluation metrics. Detailed metric definitions, training setups, including hyperparameters are provided in the Appendix. All evaluations are conducted on our processed subset of the MIMIC-IV dataset (Johnson et al. 2023), with scope constrained by computational and cost considerations.

Main Results. As shown in Table 1, **existing LLMs struggle with effective medical history taking**, which is consistent with prior studies (Johri et al. 2025; Li et al. 2024). Among them, proprietary LLMs like GPT-4o and Gemini-2.5-flash perform best, achieving top-1 diagnostic accuracies of 49.0% and 51.4%, respectively. In contrast, public models such as Qwen2.5-7B-Instruct and Qwen3-8B show much weaker performance, with F1 scores below 20% and Top-1 accuracies under 40%. DeepSeek-R1-0528-Qwen3-8B is the strongest open-source general model, reaching an F1 of 29.6% and Top-1 accuracy of 37.2%. Interestingly, models like Gemini-2.5-flash and o4-mini engage in long conversations (over 27 turns on average) but still achieve low recall, indicating a lack of ability to ask clinically relevant questions. This highlights the gap between conversational fluency and clinically meaningful reasoning.

Several domain-specific medical LLMs were also evaluated. MedGemma-4B-it and MedGemma-27B-text-it perform relatively well in diagnosis (Top-1: 40.6% and 52.8%), likely benefiting from medical pretraining. In contrast, HuatuoGPT-o1-8B performs poorly across all metrics, especially in history taking (F1: 0.2%), as it fails to ask follow-up questions and relies solely on the chief complaint. DoctorAgent-RL, despite being trained on history taking dialogues, is limited by its design to predict only a single diagnosis. It performs reasonably in information elicitation (F1: 28.4%) but fails to generalize to our setting. This is likely due to a domain mismatch, as it was trained on informal online consultations, which lack the structured, standardized symptom descriptions (e.g., onset, location, timing) found in clinical notes.

Fine-tuning significantly improves both history taking and diagnosis. Our proposed models, `Note2Chat-MT` and `Note2Chat-ST`, consistently outperform all baselines. Compared to the base model Qwen2.5-7B-Instruct, `Note2Chat-ST` achieves a 26.5-point absolute gain in F1 (from 19.6% to 46.1%), an 135.2% relative improvement, and a 31.2-point gain in Top-1 accuracy (from 38.8% to 70.0%). `Note2Chat-MT` achieves the highest recall (55.4%) but at the cost of longer dialogues (avg. 27.5 turns). In contrast, `Note2Chat-ST` delivers the best overall performance across F1, precision, and Top-K accuracy, while using fewer turns (17.3 on average). This demonstrates the effectiveness of the proposed `Note2Chat` framework in gathering clinically relevant information efficiently and making accurate differential diagnosis.

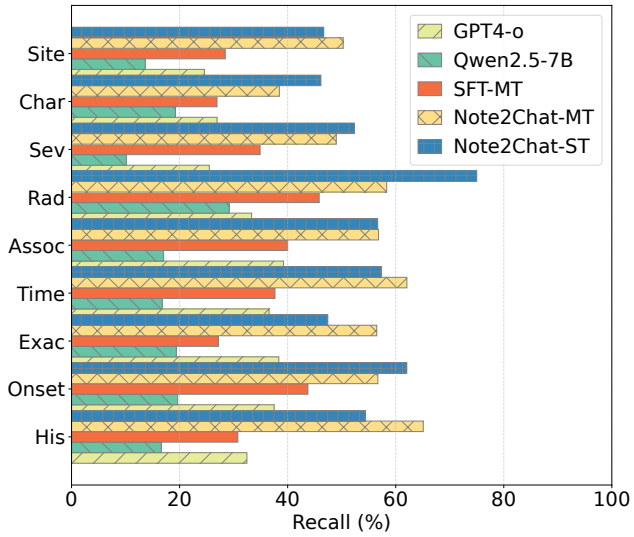

Figure 2: Recall across symptom categories.

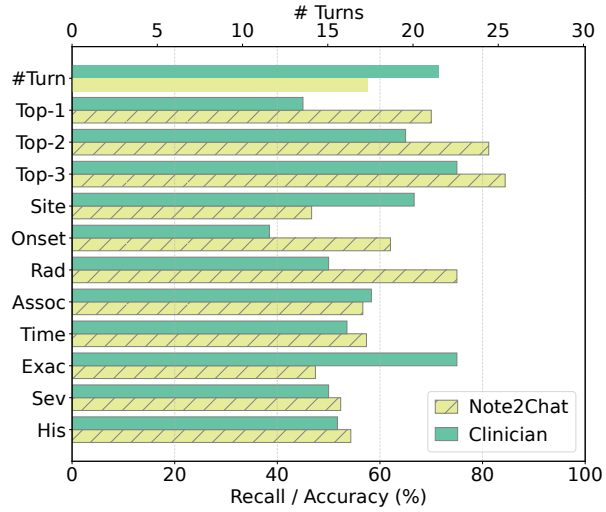

Figure 3: Comparison of history taking performance between model and clinician.

Ablation Study. We assess the contribution of each component in our `Note2Chat` framework for medical history taking. As shown in Table 2, **both multi-turn (MT) and single-turn (ST) models benefit substantially from fine-tuning.** Applying SFT alone yields notable gains over the base Qwen2.5-7B-Instruct model, improving average scores by 12–13 points. However, SFT alone only makes the model competitive with top-performing LLMs, suggesting that simply memorizing dialogue patterns is insufficient for effective history taking. Introducing self-augmented, imperfect trajectories significantly enhances recall and top-K accuracy, highlighting the value of diverse training signals. Further applying DPO consistently improves performance across both paradigms. With the full pipeline (SFT + Self-Aug + DPO), `Note2Chat-MT` achieves the highest recall (55.4%) but requires longer dialogues (27.5 turns). In contrast, `Note2Chat-ST` outperforms in all other metrics while using fewer turns (17.3), demonstrating the efficiency and effectiveness of the proposed single-turn reasoning framework for proactive and accurate history taking.

Analysis. To better understand the factors underlying effective medical history taking and how our `Note2Chat` framework improves performance, we break down recall scores by clinically meaningful symptom categories based on the SOCRATES mnemonic (Mahbubani 2023) (Site, Onset, Character, Radiation, Associated symptoms, Timing, Exacerbating/relieving factors, Severity) along with an additional History category. These categories capture essential dimensions of structured symptom characterization in clinical history taking. As shown in Figure 2, even powerful proprietary models like GPT-4o underperform on essential aspects such as Site (13.6%) and Severity (10.1%), highlighting a gap in alignment with structured clinical inquiry. SFT yields moderate gains over the base Qwen2.5-7B model across all categories. In contrast, both **`Note2Chat-MT` and `Note2Chat-ST` achieve consistently higher recall across all categories, particularly excelling in Onset, Radiation, and History.** Notably, `Note2Chat-MT` reaches the highest overall recall by leveraging longer interactions (27.5 turns on average), demonstrating its effectiveness in information gathering.

Comparison with Clinicians. Finally, we compare our method with a practicing clinician on a small held-out test set of 20 patient cases across 10 diseases, constrained by available resources. As shown in Figure 3, our model achieves comparable performance in both diagnosis accuracy and information gathering. Notably, **it mirrors the clinician's behavior in eliciting clinically meaningful symptoms.** While this limited-scale validation with simulated patients is far from conclusive, it highlights the potential of LLMs for supporting real-world history taking.

Conclusion

We have presented `Note2Chat`, a note-driven framework for training LLMs to perform clinically effective history taking and differential diagnosis. By leveraging medical notes as silver-standard supervision, our approach enables models to ask relevant follow-up questions, prioritize key findings, and determine when to conclude the conversation. Through a multi-stage training pipeline combining supervised fine-tuning, self-augmented trajectory sampling, and preference optimization, `Note2Chat` achieves substantial gains in both information elicitation and diagnostic accuracy. Our proposed single-turn reasoning paradigm enables fine-grained, verifiable supervision at each step of the dialogue. This design not only improves transparency and adaptability but also outperforms multi-turn baselines with fewer dialogue turns. Empirical results show that `Note2Chat` consistently outperforms both general-purpose and medical-domain LLMs. Additionally, clinician comparisons show promising alignment in symptom gathering, suggesting real-world applicability for AI-assisted history taking.

Acknowledgements

This research/project is supported by the National Research Foundation Singapore under the AI Singapore Programme (AISG Award No: AISG2-TC-2023-013). This research is also supported by the SingHealth Duke-NUS Academic Medical Centre (AMC) – A*STAR Healthcare Translation Partnership (HTP) grant No. I24D1AG022 and I24D1AG085.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Cabral, S.; Restrepo, D.; Kanjee, Z.; Wilson, P.; Crowe, B.; Abdulnour, R.-E.; and Rodman, A. 2024. Clinical Reasoning of a Generative Artificial Intelligence Model Compared With Physicians. *JAMA Internal Medicine*, 184(5): 581–583.

Chandra, M.; Sriraman, S.; Khanuja, H. S.; Jin, Y.; and De Choudhury, M. 2025. Reasoning Is Not All You Need: Examining LLMs for Multi-Turn Mental Health Conversations. *arXiv preprint arXiv:2505.20201*.

Chen, J.; Cai, Z.; Ji, K.; Wang, X.; Liu, W.; Wang, R.; and Wang, B. 2025a. Towards Medical Complex Reasoning with LLMs through Medical Verifiable Problems. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 14552–14573. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.

Chen, J.; Wei, Z.; Zhang, W.; Hu, Y.; and Zhang, Q. 2025b. CliniChat: A Multi-Source Knowledge-Driven Framework for Clinical Interview Dialogue Reconstruction and Evaluation. *arXiv preprint arXiv:2504.10418*.

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

Fan, Z.; Tang, J.; Chen, W.; Wang, S.; Wei, Z.; Xi, J.; Huang, F.; and Zhou, J. 2024. AI Hospital: Benchmarking large language models in a multi-agent medical interaction simulator. *arXiv preprint arXiv:2402.09742*.

Fansi Tchango, A.; Goel, R.; Wen, Z.; Martel, J.; and Ghosn, J. 2022. DDXPlus: A new dataset for automatic medical diagnosis. *Advances in neural information processing systems*, 35: 31306–31318.

Feng, Y.; Wang, J.; Zhou, L.; Lei, Z.; and Li, Y. 2025. DoctorAgent-RL: A Multi-Agent Collaborative Reinforcement Learning System for Multi-Turn Clinical Dialogue. *arXiv preprint arXiv:2505.19630*.

Gatto, J.; Seegmiller, P.; Burdick, T. E.; Khayal, I. S.; DeLozier, S.; and Preum, S. M. 2025. Follow-up Question Generation For Enhanced Patient-Provider Conversations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 25222–25240. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Goh, E.; Gallo, R.; Hom, J.; Strong, E.; Weng, Y.; Kerman, H.; Cool, J. A.; Kanjee, Z.; Parsons, A. S.; Ahuja, N.; et al. 2024. Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. *JAMA Network Open*, 7(10): e2440969.

Guo, D.; Yang, D.; Zhang, H.; et al. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081): 633–638.

Guyatt, G.; Rennie, D.; Meade, M. O.; and Cook, D. J. 2015. *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice*. New York: McGraw-Hill Education, 3 edition. ISBN 978-0-07-1790710.

Hager, P.; Jungmann, F.; Holland, R.; Bhagat, K.; Hubrecht, I.; Knauer, M.; Vielhauer, J.; Makowski, M.; Braren, R.; Kaissis, G.; and Rueckert, D. 2024. Evaluation and Mitigation of the Limitations of Large Language Models in Clinical Decision-Making. *Nature Medicine*, 30(9): 2613–2622.

Henderson, M.; Tierney, L. M.; and Smetana, G. W. 2012. *The Patient History: EvidenceBased Approach*. New York, NY: McGrawHill Education / Medical, 2nd edition. ISBN 9780071624947.

Johnson, A. E.; Bulgarelli, L.; Shen, L.; Gayles, A.; Shammout, A.; Horng, S.; Pollard, T. J.; Hao, S.; Moody, B.; Gow, B.; et al. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data*, 10(1): 1.

Johri, S.; Jeong, J.; Tran, B. A.; Schlessinger, D. I.; Wongvibulsin, S.; Barnes, L. A.; Zhou, H.-Y.; Cai, Z. R.; Van Allen, E. M.; Kim, D.; Daneshjou, R.; and Rajpurkar, P. 2025. An Evaluation Framework for Clinical Use of Large Language Models in Patient Interaction Tasks. *Nature Medicine*, 31(1): 1–10.

Kuriakose, T. 2020. History Taking: The Most Important Clinical Test. In *Clinical Insights and Examination Techniques in Ophthalmology*, 21–29. Singapore: Springer Nature Singapore Pte Ltd, 1 edition. ISBN 978-981-15-2889-7.

Labrak, Y.; Bazoge, A.; Morin, E.; Gourraud, P.-A.; Rouvier, M.; and Dufour, R. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Lai, Y.; Zhong, J.; Li, M.; Zhao, S.; and Yang, X. 2025. Med-R1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv preprint arXiv:2503.13939*.

Li, S. S.; Balachandran, V.; Feng, S.; Ilgen, J. S.; Pierson, E.; Koh, P. W.; and Tsvetkov, Y. 2024. MediQ: Question-Asking LLMs and a Benchmark for Reliable Interactive Clinical Reasoning. In *Advances in Neural Information Processing Systems 37*.

Liu, L.; Yang, X.; Li, F.; Chi, C.; Shen, Y.; Lyu, S.; Zhang, M.; Ma, X.; Lv, X.; Ma, L.; Zhang, Z.; Xue, W.; Huang, Y.; and Gu, J. 2024. Towards Automatic Evaluation for LLMs'

Clinical Capabilities: Metric, Data, and Algorithm. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5466–5475. ACM.

Liu, X.; Sun, D.; Fung, Y. R.; Hakkani-Tür, D.; and Abdelzaher, T. 2025. DocCHA: Towards LLM-Augmented Interactive Online diagnosis System. *arXiv preprint arXiv:2507.07870*.

Mahbubani, K. 2023. *Basic History Taking*, 1–5. Cham: Springer International Publishing. ISBN 978-3-031-29897-4.

McDuff, D.; Schaekermann, M.; Tu, T.; Palepu, A.; Wang, A.; Garrison, J.; Singhal, K.; Sharma, Y.; Azizi, S.; Kulkarni, K.; et al. 2025. Towards Accurate Differential Diagnosis with Large Language Models. *Nature*, 642: 451–457.

Nierenberg, R. J. 2020. Using the chief complaint driven medical history: theoretical background and practical steps for student clinicians. *MedEdPublish*, 9: 17.

Nori, H.; Daswani, M.; Kelly, C.; Lundberg, S.; Ribeiro, M. T.; Wilson, M.; Liu, X.; Sounderajah, V.; Carlson, J.; Lungren, M. P.; Gross, B.; Hames, P.; Suleyman, M.; King, D.; and Horvitz, E. 2025. Sequential Diagnosis with Language Models. *arXiv preprint arXiv:2506.22405*.

Nori, H.; King, N.; McKinney, S. M.; Carignan, D.; and Horvitz, E. 2023a. Capabilities of GPT-4 on Medical Challenge Problems. *arXiv preprint arXiv:2303.13375*.

Nori, H.; Lee, Y. T.; Zhang, S.; Carignan, D.; Edgar, R.; Fusi, N.; King, N.; Larson, J.; Li, Y.; Liu, W.; et al. 2023b. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. *arXiv preprint arXiv:2311.16452*.

Nori, H.; Usuyama, N.; King, N.; McKinney, S. M.; Fernandes, X.; Zhang, S.; and Horvitz, E. 2024. From Medprompt to o1: Exploration of Run-Time Strategies for Medical Challenge Problems and Beyond. *arXiv preprint arXiv:2411.03590*.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36: 53728–53741.

Rose, D. P.; Hung, C.-C.; Lepri, M.; Alqassem, I.; Gashteovski, K.; and Lawrence, C. 2025. MEDDxAgent: A Unified Modular Agent Framework for Explainable Automatic Differential Diagnosis. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13803–13826. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Saab, K.; Freyberg, J.; Park, C.; Strother, T.; Cheng, Y.; Weng, W.-H.; Barrett, D. G.; Stutz, D.; Tomasev, N.; Palepu, A.; et al. 2025. Advancing Conversational Diagnostic AI with Multimodal Reasoning. *arXiv preprint arXiv:2505.04653*.

Saab, K.; Tu, T.; Weng, W.-H.; Tanno, R.; Stutz, D.; Wulczyn, E.; Zhang, F.; Strother, T.; Park, C.; Vedadi, E.; et al. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Saley, V. V.; Saha, G.; Das, R. J.; Raghu, D.; and ., M. 2024. MediTOD: An English Dialogue Dataset for Medical History Taking with Comprehensive Annotations. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 16843–16877. Miami, Florida, USA: Association for Computational Linguistics.

Schmidgall, S.; Ziaei, R.; Harris, C.; Reis, E.; Jopling, J.; and Moor, M. 2024. AgentClinic: a Multimodal Agent Benchmark to Evaluate AI in Simulated Clinical Environments. *arXiv preprint arXiv:2405.07960*.

Sellergren, A.; Kazemzadeh, S.; Jaroensri, T.; Kiraly, A.; Traverse, M.; Kohlberger, T.; Xu, S.; Jamil, F.; Hughes, C.; Lau, C.; et al. 2025. MedGemma Technical Report. *arXiv preprint arXiv:2507.05201*.

Sun, Z.; Liu, Z.; Luo, C.; Chu, J.; and Huang, Z. 2025. Improving interactive diagnostic ability of a large language model agent through clinical experience learning. *arXiv preprint arXiv:2503.16463*.

Team, Q. 2024. Qwen2.5: A Party of Foundation Models. https://qwenlm.github.io/blog/qwen2.5/. Accessed: 2024-09-19.

Team, Q. 2025. Qwen3 Technical Report. arXiv:2505.09388.

Tu, T.; Schaekermann, M.; Palepu, A.; Saab, K.; Freyberg, J.; Tanno, R.; Wang, A.; Li, B.; Amin, M.; Cheng, Y.; et al. 2025. Towards conversational diagnostic artificial intelligence. *Nature*, 642(8067): 442–450.

Wang, Z.; Li, H.; Huang, D.; Kim, H.-S.; Shin, C.-W.; and Rahmani, A. M. 2025. HealthQ: Unveiling questioning capabilities of llm chains in healthcare conversations. *Smart Health*, 100570.