Inflated Excellence or True Performance? Rethinking Medical Diagnostic Benchmarks with Dynamic Evaluation

Anonymous ACL submission

Abstract

Medical diagnostics is a high-stakes and complex domain that is critical to patient care. However, current evaluations of large language models (LLMs) are fundamentally misaligned with real-world clinical practice. Most rely on benchmarks derived from public exams, raising contamination bias that can inflate performance, and they overlook the confounded nature of real consultations beyond textbook cases. Recent dynamic evaluations offer a promising alternative, but often remain insufficient for realistic diagnostic benchmarking, with limited coverage of clinically grounded confounders and trustworthiness beyond accuracy. To address these gaps, we propose **DyReMe**, a dynamic benchmark for medical diagnostics that better reflects real clinical practice. Unlike static exam-style questions, **DyReMe** generates fresh, consultation-like cases that introduce distractors such as differential diagnoses and common misdiagnosis factors. It also varies expression styles to mimic diverse real-world query habits. Beyond accuracy, **DyReMe** evaluates LLMs on three additional clinically relevant dimensions: veracity, helpfulness, and consistency. Our experiments demonstrate that this dynamic approach yields more challenging and realistic assessments, revealing significant misalignments between the performance of state-of-the-art LLMs and real clinical practice. These findings highlight the urgent need for evaluation frameworks that better reflect the demands of trustworthy medical diagnostics.

1 Introduction

Accurate medical diagnostics is vital for patient health and effective treatment. However, the inherent complexity (e.g., the diversity of symptom-diagnosis relationships) and external factors (e.g., incomplete information and various expression styles) pose significant challenges (Studdert et al., 2005; Naik et al., 2022). Misdiagnoses can lead to severe consequences, such as higher mortality

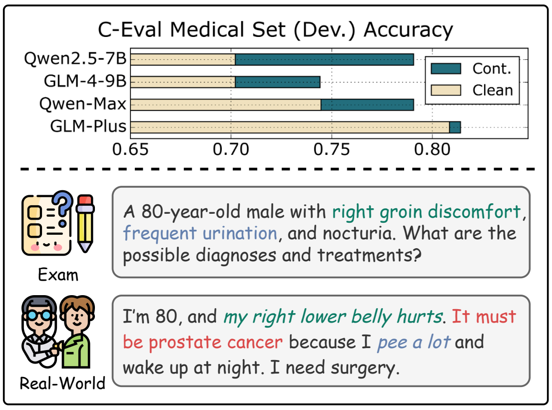


Figure 1: (Top) Contamination data experiments on C-Eval (Huang et al., 2023; Li et al., 2024a). (Bottom) Unlike standardized patients in current benchmarks, real-world patients express symptoms unprofessionally and may misdiagnose themselves, misleading doctors.

rates (Pourafkari et al., 2017) and increased healthcare burdens (Juarez-Garcia et al., 2006), underscoring the need for trustworthy diagnostic tools aligned with clinical practice (Pai et al., 2024).

Recent advances in large language models (LLMs) show promise in assisting healthcare professionals by processing medical knowledge (Chen et al., 2024b) and supporting diagnostic decisions (Chen et al., 2024a). These LLMs can analyze clinical cases (Wang et al., 2024b), identify patterns in laboratory medicine data (Ullah et al., 2024), and potentially improve diagnostic accuracy (Fan et al., 2024). To assess their capabilities, *static* benchmarks based on medical exams have been developed (Liu et al., 2023; Cai et al., 2024), where the test questions remain fixed across models and time (Zhu et al., 2024; Wang et al., 2025).

However, a key question remains: *Do static benchmarks reflect real-world diagnostic capabilities of LLMs?* (Jiang et al., 2024; Gupta et al., 2024). Two limitations are particularly relevant: 1) Biased estimation of real-world capability: Because many medical benchmarks are *public and static*, they are vulnerable to *data contamination* (Li et al., 2024b; Xu et al., 2024). This overlap biases evaluation: high scores may reflect exposure rather than generalizable reasoning, overestimating real-world performance. 2) Misalignment with real-world scenarios (Liu et al., 2024b; Park et al., 2024): Even when contamination is controlled, exam-style benchmarks adopt standardized, well-formed case descriptions and an accuracy-centric protocol, whereas real-world queries are often incomplete, lay, and confounded by self-diagnosis, which can mislead clinical decision-making (Graber et al., 2005; Norman et al., 2017). Fig.1 further illustrates this gap: the top panel shows contamination-induced inflation, while the bottom panel highlights patient-style expressions and diagnostic confounders absent from exam questions. These issues call for an evaluation approach that mitigates contamination and better aligns with real-world diagnostic interactions, without relying on costly annotation (Lyu et al., 2024).

One natural direction is *dynamic* evaluation, which generates new test cases by transforming existing questions (e.g., rewriting (Wang et al., 2025), adding noise (Zhu et al., 2024), or altering graphs (Zhu et al., 2023)) to reduce direct overlap with public benchmarks. However, such transformations are often surface-level and largely preserve the underlying clinical setting, leaving real-world misalignment unresolved in diagnostic confounders (Graber et al., 2005) and patient-style expressions. Moreover, it still does not change the sole focus on accuracy. Consequently, dynamic evaluation remains insufficient for benchmarking LLMs under realistic diagnostic conditions. This motivates an evaluation framework that (i) generates *unseen* cases *with realistic confounders* and (ii) evaluates *trustworthiness beyond accuracy*.

To address both contamination-induced evaluation bias and real-world misalignment, we propose **DyReMe**, a **Dy**namic, **Re**alistic evaluation framework for **Me**dical diagnostics. **DyReMe** has two key components: `DyGen` and `EvalMed`. `DyGen` generates realistic diagnostic benchmarks by creating *unseen* questions that incorporate differential diagnoses and misdiagnosis factors as diagnostic distractors (e.g., a *sinusitis* patient self-diagnosed with *periodontitis*). Additionally, `DyGen` takes patient expression features into account, ensuring that the generated questions reflect real-world query styles (e.g., a patient describing *radiating pain* as "*an electric shock running down my arm*"). Together, these designs explicitly inject real-world confounders and patient-style expressions, reducing the misalignment between benchmarks and practical diagnostic interactions. `EvalMed` assesses four aspects: accuracy, veracity, helpfulness, and consistency. It checks whether the model gives the correct diagnosis, corrects health rumors, provides useful explanations, and gives stable answers. By moving beyond an accuracy-only protocol, `EvalMed` addresses evaluation-protocol misalignment and reduces over-optimistic conclusions drawn from static, exam-style benchmarks. Experiments demonstrate that **DyReMe** not only generates unseen, confounder-rich diagnostic cases (`DyGen`) but also provides trustworthy, human-aligned evaluations (`EvalMed`), revealing limitations of LLMs under realistic diagnostic conditions.

Contributions. (1) We propose **DyReMe**, a dynamic evaluation framework for medical diagnostics that combines realistic benchmark generation (`DyGen`) with trustworthy evaluation (`EvalMed`). (2) Using `DyGen`, we construct a challenging benchmark that reflects key clinical confounders. (3) We evaluate 12 LLMs with `EvalMed`, uncovering limitations in their diagnostic capabilities.

2 Related Work

Static Benchmarks for Medical LLMs. Medical LLMs are often evaluated on benchmarks such as PubmedQA (Jin et al., 2019), MedQA (Jin et al., 2021), CMB (Wang et al., 2024b), and CMExam (Liu et al., 2023), which come from static medical exams. Although models like GPT-4o (Hurst et al., 2024) perform well on such benchmarks, their public availability raises concerns about contamination and inflated scores. Moreover, exam-style questions often fail to capture real clinical complexity. Newer datasets such as MedBench (Cai et al., 2024), DxBench (Chen et al., 2024a), CliMedBench (Ouyang et al., 2024), and others (Ness et al., 2024; Kim et al., 2024; Zhang et al., 2025; Bedi and et al., 2025; Pan et al., 2025) incorporate expert annotations and better reflect clinical reasoning (e.g., RJUA (Lyu et al., 2024)), but remain costly, prone to contamination, and limited in representativeness. Thus, despite higher reasoning fidelity, they do not scale or refresh fast enough to match the pace of medical LLMs.

Dynamic Evaluation. Dynamic evaluation has been proposed to address the limitations of static

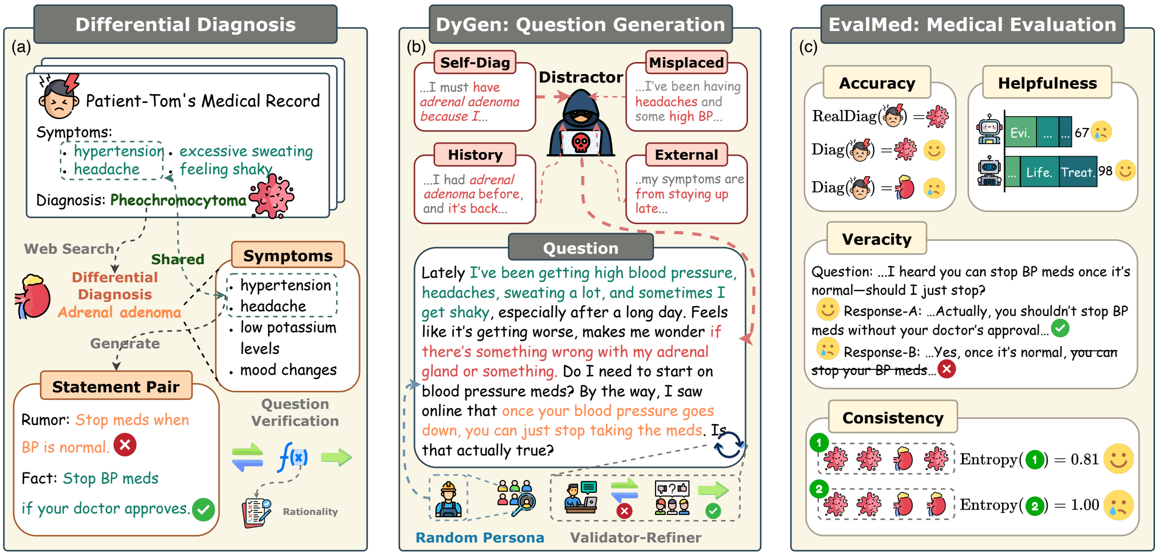


Figure 2: Overview of **DyReMe**. (a) Differential diagnosis construction and medical rumor generation. (b) Question generation with trap selection, persona style, and refinement. We focus on **Chinese** questions and the example question is translated from Chinese. (c) `EvalMed` assesses Accuracy, Veracity, Helpfulness, and Consistency.

benchmarks by rewriting, perturbing, or paraphrasing raw questions (Kiela et al., 2021; Zhang et al., 2024; White et al., 2024; Wang et al., 2024a) . In this work, we focus on generating new medical diagnostic benchmarks from seed cases. Among these approaches, dynamic evaluation methods based on LLMs, such as DyVal2 (Zhu et al., 2024) and Self-Evolving (Wang et al., 2025) employ LLMs as generators to improve scalability and adaptability. DyVal2 introduces "probing" and "judging" agents to create new questions through paraphrasing, adding noise, and permutation. Self-Evolving applies six different reframing operations to construct evolving instances, testing LLMs across query variations, noise, and problem-solving robustness. However, these methods overlook the specific requirements of medical diagnostics and fail to address the limitations of static benchmarks, as they mainly rely on superficial transformations.

Summary. Existing benchmarks overestimate performance and fail to reflect clinical practice. Dynamic methods attempt to bridge this gap but still miss real-world distractors and trustworthy assessment. To address this gap, we introduce **DyReMe**.

3 Methodology

We propose **DyReMe** (Fig.2), a dynamic evaluation framework that better aligns with real-world clinical practice. It comprises two components: (1) `DyGen`, a generation module that creates realistic and challenging questions, and (2) `EvalMed`, an evaluation module that assesses LLMs across four clinically relevant dimensions: *accuracy*, *veracity*, *helpfulness*, and *consistency*. We begin by formally defining the diagnostic task. Given a question q based on a set of symptoms $S = \{s_i\}_{i=1}^{N}$ and an underlying diagnosis d, the LLM M, parameterized by θ, generates a response $\hat{a} = M(q)$. A scoring function f is used to assess the quality of $\hat{a}$.

3.1 `DyGen`: Dynamic Generation

Current diagnostic benchmarks oversimplify the complexity and variability of real-world diagnostics, resulting in a misalignment with actual clinical consultations. Trustworthy diagnosis requires consideration of differential diagnoses (Arter and Jenkins, 1979) and misdiagnosis factors (Graber et al., 2005; Norman et al., 2017), collectively referred to as "diagnostic distractors." To simulate real-world complexity, `DyGen` first constructs diagnostic distractors by integrating differential diagnoses and misdiagnosis factors. It then incorporates these distractors into the questions and rephrases them to match authentic clinical expressions. Finally, `DyGen` refines the questions using a validator-refiner iterative loop to ensure clinical validity and maintain real-world complexity.

Differential Diagnosis. `DyGen` retrieves similar diagnoses d_{dis} for a given original diagnosis d_{org}, mimicking real-world differential diagnosis. This is achieved using Retrieval-Augmented Generation ($\mathcal{G}_{\text{RAG}}$) (Lewis et al., 2020). For example, $\mathcal{G}_{\text{RAG}}$ retrieves d_{dis} = "*Adrenal adenoma*" as a differential diagnosis for d_{org} = "*Pheochromocytoma*" based on medical encyclopedias, since both share common symptoms like *hypertension* and *headache* (Fig.2a). Formally, $d_{\text{dis}} = \mathcal{G}_{\text{RAG}}(d_{\text{org}})$.

Misdiagnosis Factors. Anchor bias (Graber et al., 2005), posterior probability error (Nendaz and Perrier, 2012), distraction (Graber et al., 2005), and symptom overestimation (Braun et al., 2017) pose challenges in clinical practice. To reflect these factors, DyGen designs four diagnostic traps (Fig.2b): $\mathcal{S}$={self-diagnosis, distracting history, external noise, misplaced symptoms}. Given a question q_{org} (e.g., "*What's wrong with me if I have hypertension and headache?*"), DyGen selects a trap s and a differential diagnosis d_{dis} (e.g., "*Adrenal adenoma*"), and combines them to form a misleading question q_{trap} (e.g., "*I might have an adrenal adenoma due to hypertension and headache. Can you give me medicine?*"):

$$q_{\text{trap}} = \mathcal{T}_{\text{trap}}(q_{\text{org}}, s, d_{\text{dis}}), \quad s \sim \mathcal{U}(\mathcal{S}). \quad (1)$$

Here, $\mathcal{T}_{\text{trap}}$ denotes the operation of constructing and adding distractors into questions, and $\mathcal{U}(\cdot)$ is the uniform distribution function. These traps simulate common real-world diagnostic pitfalls, making the questions more realistic, challenging, and better aligned with actual clinical scenarios. See Appendix A for detailed descriptions.

Expression Style. Patients often describe symptoms in lay language (Zeng and Tse, 2006; Forbush et al., 2013), reflecting their subjective perceptions rather than formal clinical nomenclature. To capture this, DyGen employs a persona-driven stylistic adaptation mechanism that models diverse patient expression styles (Ge et al., 2024). However, integrating persona information into questions may introduce unintended causal confounders and compromise the validity of the ground truth. For example, personas (e.g., *miners*) may correlate with specific diagnoses (e.g., *pneumoconiosis*). To avoid this, DyGen applies an indirect adaptation operation ($\mathcal{T}_{\text{persona}}$). It first extracts expression features (e.g., knowledge level, clarity, and communication style) from a persona b, and then uses them to rephrase the question. This process is formalized as:

$$q_{per} = \mathcal{T}_{\text{persona}}(q_{\text{trap}}, b), \quad b \sim \mathcal{U}(\mathcal{B}). \quad (2)$$

For example, q_{per} is "*My blood pressure's high and I keep getting bad headaches, and I think there's a bump on my kidney. Can you give me something for it?*" when b = *"Mason"*, a persona with limited medical knowledge. This variation mirrors real-world patient communication and introduces challenges that require LLMs to align diagnostic reasoning with non-standardized narratives.

Validator–Refiner Iterative Loop. Inspired by critique-and-revision paradigms (Tan et al., 2023; Gou et al., 2024), we implement an iterative refinement loop to ensure the quality and realism of generated diagnostic questions. At iteration t, a validator $\mathcal{V}$ evaluates candidate question q_t along four dimensions (challenge, logical consistency, symptom accuracy, and trap effectiveness). If q_t passes, the process terminates with $q^* = q_t$. Otherwise, q_t is returned to a refiner $\mathcal{R}$, which revises the question based on validator feedback $\mathcal{F}(q_t, \mathcal{V})$:

$$\begin{cases} q_t, & \mathcal{V}(q_t) = 1 \\ \mathcal{R}(q_t, \mathcal{F}(q_t, \mathcal{V})), & \text{otherwise.} \end{cases} \quad (3)$$

This loop improves question quality until all constraints are satisfied. This process yields clinically realistic and diagnostically rigorous questions.

3.2 EvalMed: Medical Evaluation

To assess diagnostic performance beyond *accuracy*, EvalMed also measures *veracity*, *helpfulness*, and *consistency*, capturing whether responses are truthful, useful, and stable across scenarios.

Veracity. Veracity assesses whether LLMs can identify and correct medical misinformation, helping prevent *infodemic* (Zarocostas, 2020; Orso et al., 2020). Existing benchmarks validate outputs against external knowledge bases (Dmonte et al., 2024; Song et al., 2024), but this often fails due to limited coverage and does not test whether models can proactively recognize or correct false claims (Min et al., 2023). In contrast, our method tests the model's ability to rectify medical rumors by presenting it with generated false statements. Inspired by SimpleQA (Wei et al., 2024), we use $\mathcal{G}_{\text{RAG}}$ to generate rumor-fact pairs $RF(m) = (e^m_{\text{rumor}}, e^m_{\text{fact}})$ for medical entity m. Given m = *"High BP"*, we might have e^m_{rumor} = *"High BP affects the bones"* and e^m_{fact} = *"High BP affects the heart"*. We retain only valid pairs, i.e., those satisfying a rationality check $\rho(\cdot) = 1$, formalized as:

$$RF_{\text{valid}}(m) = \{\, r \in RF(m) \mid \rho(r) = 1 \,\}. \quad (4)$$

To evaluate veracity, DyGen inserts a rumor into each question (e.g., "*My BP is up, I heard that high BP affects the bones. Can you recommend some medications for bones?*"). EvalMed then tests whether the LLM rectifies the rumor:

$$\text{Ver}(M) = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbb{I}_r(q, \hat{a}), \quad (5)$$

where $\text{Ver}(M)$ is the score and $\mathcal{Q}$ is the set of all test questions. $\mathbb{I}_r(\cdot,\cdot)$ is the indicator function (implemented by a worker LLM) that determines whether the LLM's response rectifies the rumor.

Helpfulness. LLMs evade responsibility by providing vague or evasive answers. Thus, we evaluate *helpfulness* (Yang et al., 2019; Shen et al., 2024) by measuring which criteria the response meets. Helpfulness is essential in medical diagnostics, where answers must be clear, actionable, and in line with professional standards (Larasati et al., 2023; Luo et al., 2024). We define three helpfulness criteria based on real-world medical platform guidelines (DingXiang, 2022): *diagnosis evidence*, *treatment suggestions*, *lifestyle suggestions*. For each diagnosis in the benchmark, a knowledge base is built using $\mathcal{G}_{\text{RAG}}$, sourcing authoritative information from medical encyclopedias. For each test question, EvalMed retrieves relevant context and assigns a score k_h^q for each criterion h. It then checks whether the response covers these points. For example, "*You may have pheochromocytoma due to high BP and headache. Pheochromocytoma surgical excision could be a good option*" scores highly, as it includes both evidence and treatment suggestions. Formally:

$$\text{Help}(M) = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \sum_{h \in \mathcal{H}} w_h \Phi_h\left(\hat{a}, k_h^q\right), \quad (6)$$

where $\Phi_h(\cdot,\cdot)$ measures coverage of the score-points for criterion h, and w_h is its weight.

Consistency. Consistency evaluates the stability of model predictions by computing the entropy of diagnoses across different variants of the same case. In high-stakes domains like medical diagnostics (Kadavath et al., 2022; Yadkori et al., 2024), inconsistent answers can erode user trust (Wu et al., 2024). To quantify consistency, we draw inspiration from *Semantic Entropy* (Farquhar et al., 2024). To reduce superficial variations, diagnostic terms are normalized. Consistency is then calculated as the normalized *Information Entropy* (Shannon, 1948) of the model's diagnosis distribution for each prediction group (i.e., all m variants of a original case). The score is mapped to $[0, 100]$, where a higher value means more consistent predictions. As shown in the lower-right corner of Fig.2c, Prediction Group-① is more consistent than Prediction Group-② due to more concentrated predictions. The process is formalized as follows:

$$\text{Cons}(M) = \frac{1}{|\mathcal{P}|} \sum_{p_i \in \mathcal{P}} \left(1 - \frac{E_{p_i}}{\log m}\right), \quad (7)$$

where E_{p_i} is the entropy of diagnoses in prediction group p_i, and $\mathcal{P}$ is the set of all prediction groups (one per original case). A higher consistency score indicates more stable predictions across different paraphrasings or expression styles.

In summary, DyGen generates challenging questions with real-world distractors and diverse styles, and EvalMed evaluates them across accuracy, veracity, helpfulness, and consistency, enabling **DyReMe** to better match practical clinical needs.

4 Experiments

This section investigates the following research questions: (**RQ1**) Does DyGen generate high-quality data? (**RQ2**) Are the evaluation results from EvalMed reliable? (**RQ3**) Are current LLMs trustworthy when conducting diagnostics?

4.1 Experimental Setup

To evaluate **DyReMe**, we compare it against static and dynamic baselines. We build a diagnostic benchmark with DyGen by seeding from three static datasets: DxBench (Chen et al., 2024a), DDXPlus (Tchango et al., 2022), and Dxy (Xu et al., 2019). DxBench covers 461 diseases; DDXPlus is a synthetic differential-diagnosis dataset; Dxy is derived from real-world consultations. For dynamic evaluation, we include DyVal2 (Zhu et al., 2024) and Self-Evolving (Wang et al., 2025). All methods use GPT-4.1 (OpenAI, 2025b) as the generator (generation temperature 0.7; verification temperature 0). Our reliability checks show no measurable impact from generator self-recognition or evaluator choice (Appendix E). Nevertheless, we conservatively exclude GPT-4.1 from evaluation to avoid any potential self-recognition effects (Davidson

Model	Static				Dynamic		
	DDXPlus	DxBench	Dxy	Avg.	Self-Evolv.$_\triangle$	DyVal2$_\triangle$	DyReMe$_\triangle$
GPT-5† (2025a)	82.00	71.69	72.82	73.76	$74.29_{+0.72\%}$	<u>70.73</u>$_{-4.11\%}$	**67.67**$_{-8.25\%}$
DeepSeek-V3† (2024a)	80.78	70.50	77.02	72.92	$73.13_{+0.29\%}$	<u>69.50</u>$_{-4.69\%}$	**65.26**$_{-10.51\%}$
GPT-4o† (2024)	81.11	70.15	74.11	72.53	$72.98_{+0.62\%}$	<u>69.67</u>$_{-3.94\%}$	**64.74**$_{-10.75\%}$
GPT-4o-mini† (2024)	72.22	<u>66.08</u>	73.46	67.76	$70.26_{+3.68\%}$	$66.56_{-1.78\%}$	**62.35**$_{-7.99\%}$
MedGemma-27B💊 (2025)	76.78	68.24	78.32	70.56	$71.48_{+1.30\%}$	<u>67.70</u>$_{-4.06\%}$	**62.97**$_{-10.76\%}$
WiNGPT2-9B💊 (2025)	68.96	65.93	70.89	66.85	$67.30_{+0.68\%}$	<u>62.30</u>$_{-6.81\%}$	**59.89**$_{-10.41\%}$
Qwen3-32B (2025)	76.67	<u>67.16</u>	77.02	73.62	$71.55_{+2.71\%}$	$68.28_{-1.98\%}$	**63.85**$_{-8.34\%}$
Gemma-3-27B (2025)	74.78	67.48	72.81	69.25	$69.37_{+0.18\%}$	<u>66.04</u>$_{-4.63\%}$	**61.94**$_{-10.55\%}$
GLM-4-32B (2024)	72.22	68.67	73.14	69.66	$69.85_{+0.27\%}$	<u>67.93</u>$_{-2.47\%}$	**61.96**$_{-11.05\%}$
Qwen2.5-32B (2024)	70.56	66.61	74.76	67.92	$69.35_{+2.11\%}$	<u>66.30</u>$_{-2.38\%}$	**60.05**$_{-11.59\%}$
Qwen2.5-7B (2024)	67.42	67.07	77.67	67.85	$67.07_{-1.15\%}$	<u>65.25</u>$_{-3.82\%}$	**57.86**$_{-14.71\%}$

Table 1: Diagnostic accuracy (average of Top-1, 3, and 5) is reported on static and dynamic benchmarks, with lower being harder. To compare static and dynamic methods, we use the weighted average accuracy on the three static datasets as the static baseline. Δ denotes the relative change to the static average. The best and runner-up results are **boldfaced** and <u>underlined</u>, respectively. The symbol † indicates commercial LLMs, and 💊 denotes medical LLMs. **DyReMe significantly** outperforms all baselines ($p < 0.001$, t-test with 10 runs 80% bootstrap sampling).

et al., 2024). To benchmark 12 LLMs, we further scale to 3,200 questions from 800 DxBench cases. We evaluate *accuracy*, *veracity*, *helpfulness*, and *consistency*. $\mathcal{G}_{\text{RAG}}$ uses the Volcano Engine web search API and the Douyin Encyclopedia (VolcEngine, 2025b). See Appendix B.1 for details.

4.2 Assessment of Question Quality

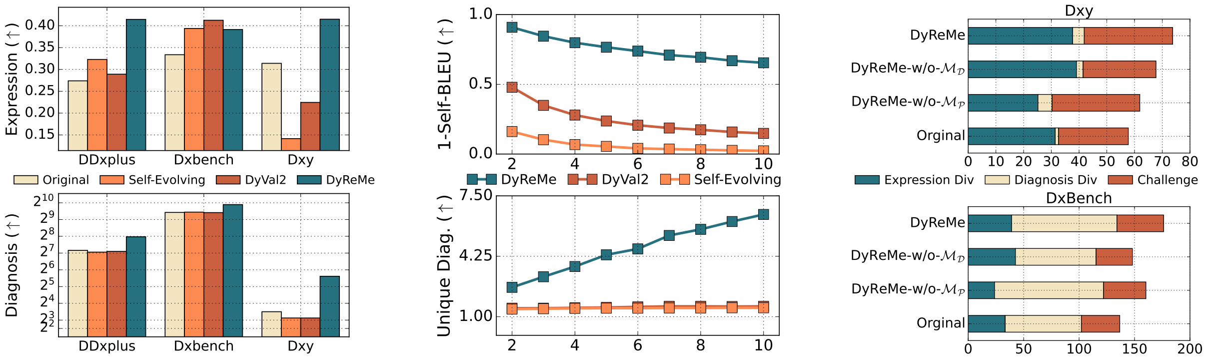
We assess question quality (**RQ1**) along *challenge* and *diversity*, ensuring the benchmark mitigates saturation and reflects real-world scenarios.

Question Challenge. We assess challenge by diagnostic accuracy across 11 LLMs (Tab. 1, details in Appendix B.2). Overall, dynamic evaluations are more challenging than static baselines, except for Self-Evolving. For example, DeepSeek-V3 drops from 72.92 (static avg.) to 69.50 on DyVal2 and 65.26 on **DyReMe**. Meanwhile, challenge is not guaranteed for all dynamic methods: when disturbances are weak, models can score higher than on static benchmarks (e.g., GPT-4o-mini achieves 70.26 on Self-Evolving and 66.56 on DyVal2 vs. 66.08 on DxBench). In contrast, **DyReMe** induces larger drops, even for GPT-5 (stronger than the generator): accuracy falls from 73.76 to 67.67, more than twice the drop on DyVal2 (-4.11%). This shows our benchmark remains challenging for frontier LLMs. By injecting real-world misdiagnosis patterns and patient-specific styles, **DyReMe** provides a more clinically grounded and rigorous test of diagnostic performance.

Question Diversity. A diagnostic benchmark should reflect the diverse ways describing symptoms and the broad range of diseases encountered in practice. Therefore, we assess diversity with two metrics: *expression diversity* (D_{exp}), calculated as the entropy of expression style distribution, and *diagnosis diversity* (D_{diag}), defined as the number of unique diagnoses (Appendix B.3). As shown in Fig. 3a, **DyReMe** improves both metrics compared to static and dynamic baselines. By comparison, other baselines show diversity levels that are similar to or even lower than those of static benchmarks. For example, both dynamic methods achieve similar diagnosis diversity as the static benchmarks. On Dxy, both dynamic methods exhibit lower expression diversity, likely due to conducting superficial transformations. By integrating expression styles, **DyReMe** notably increases expression diversity. Furthermore, by incorporating differential diagnoses, **DyReMe** nearly quadruples the number of unique diagnoses on Dxy and doubles it on other datasets. These findings suggest that **DyReMe** better represents the complexity and variability of real-world clinical cases.

4.3 Further Analysis

Extensibility Comparison. To evaluate the extensibility of **DyReMe** against existing methods, we select 100 seed samples from DxBench and generate k questions for each seed. We assess extensibility by tracking how diversity changes as k increases. As shown in Fig.3b, Self-Evolving and DyVal2 yield low initial $1-$ Self-BLEU scores,



(a) Diversity Assessment (b) Extensibility Comparison (c) Ablation Study

Figure 3: (a) Expression and diagnosis diversity. To disentangle effects of question count from diversity, we use each dataset as a seed pool and derive a same-size benchmark to ensure fair comparison between static and dynamic methods. (b) The group-level diversity changes with the increase of k. $1 -$ Self-BLEU is computed on a group of k questions, where a higher value indicates greater diversity. (c) Performance on challenge and diversity.

which quickly drop to near zero as k increases. This reveals their limited extensibility. This may result from their reliance on superficial transformations of seed questions, leading to repetitive and less diverse outputs. In contrast, **DyReMe** demonstrates a slower decline in $1 -$ Self-BLEU, maintaining higher diversity scores as k increases, which indicates better extensibility. Furthermore, the number of unique diagnoses generated by **DyReMe** increases steadily with k, further supporting its superior extensibility. These results indicate that **DyReMe** can generate significantly more diverse questions from limited seed samples, demonstrating greater extensibility than existing methods.

Ablation Study. We conduct an ablation study on two key components of DyGen—diagnostic distractors ($\mathcal{M}_\mathcal{D}$) and patient expression styles ($\mathcal{M}_\mathcal{P}$)—based on Dxy and DxBench. To assess the contribution of each component, we create two variants: **DyReMe**-w/o-$\mathcal{M}_\mathcal{D}$ and **DyReMe**-w/o-$\mathcal{M}_\mathcal{P}$. We evaluate both variants on *challenge* and *diversity* (Fig. 3c). The results show that $\mathcal{M}_\mathcal{D}$ and $\mathcal{M}_\mathcal{P}$ contribute independently, and both are essential for optimal benchmark performance. Removing $\mathcal{M}_\mathcal{P}$ leads to a marked drop in expression diversity, while removing $\mathcal{M}_\mathcal{D}$ significantly reduces diagnosis diversity and overall challenge. This underscores their respective roles in broadening expression styles and expanding disease coverage. These findings indicate that capturing both diverse patient narratives and a wide range of diagnoses is critical for diagnostic benchmark construction. Both components are crucial for ensuring that **DyReMe** reflects the variability of clinical scenarios.

4.4 Agreement with Human

We conduct a human study to assess **DyReMe**'s clinical alignment (**RQ2**), collecting 240 data points in total—a scale comparable to prior manual evaluations in medical LLM studies (Zhang et al., 2023; Wang et al., 2024b). We first verify $\mathcal{G}_{\text{RAG}}$ on retrieving differential diagnoses and helpfulness points by sampling (d_{org}, d_{dis}) pairs and score-point groups (Appendix B.5). Human annotations indicate that 86.67% of the differential-diagnosis pairs are clinically plausible (the rest remain useful distractors), and 83.37% of the score points have high agreement with authoritative sources, with the remainder rated moderate (no disagreement). We then sample questions from **DyReMe** and ask three experts to perform two tasks (Appendix B.5): (1) *Question Quality Task*: rate each question on a 1–5 scale for rationality and clinical relevance (3: moderate, 5: high); (2) *Evaluation Preference Task*: choose the more trustworthy response from a pair. We report inter-rater agreement using Gwet's AC1 (Gwet, 2008) ranging from -1 to 1. Across both tasks, the average AC1 is 0.6889, indicating strong agreement. For Task 1, DyGen questions score 3.89 on average, close to the original questions (3.97), suggesting comparable perceived quality. For Task 2, AC1 between **DyReMe** and experts reaches 0.8889, reflecting high consistency with experts. Overall, these results provide evidence that **DyReMe** aligns with real-world clinical needs.

4.5 Benchmarking Results

We benchmark 12 leading LLMs using **DyReMe** (**RQ3**). As shown in Fig.4, commercial LLMs cur-

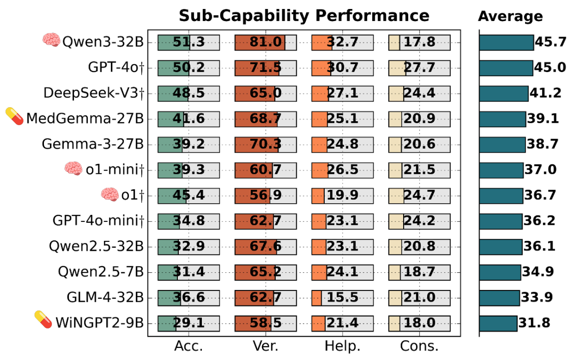


Figure 4: Results of 12 LLMs on medical diagnosis across *Accuracy*, *Veracity*, *Helpfulness*, and *Consistency*. All results are averaged over 10 runs with 80% bootstrap sampling (details in Appendix B.4). Symbol † represents commercial LLMs. Icon 💊 denotes medical LLMs and 🧠 indicates reasoning LLMs (Jaech et al., 2024).

rently maintain an overall lead, but research models are quickly catching up. For example, Qwen3-32B (45.7) and MedGemma-27B (39.1) outperform several commercial competitors like o1 and o1-mini (Jaech et al., 2024). Notably, domain-specific tuning is not a cure-all: the medically adapted WiNGPT2-9B achieves the lowest score (31.8), indicating that current adaptations may capture medical facts but often fail to handle real-world distractors and diverse expression styles. Commercial reasoning models (o1 and o1-mini) also show only moderate performance (37.0 and 36.7, respectively), likely because their training emphasizes producing a single correct answer rather than addressing health rumors or providing actionable information. All models exhibit substantial room for improvement. Even the best models, such as Qwen3-32B (51.3) and GPT-4o (50.3), still struggle with realistic diagnostic noise. Across all models, 20–40% of health rumors remain unaddressed, even for top-performing systems, posing a real risk of misinformation propagation. Most evidence and suggestions remain shallow, reflecting the difficulty current LLMs face in delivering explainable diagnoses and actionable advice. Consistency is consistently low across models, making them vulnerable to changes in input context. Overall, results from **DyReMe** indicate that current static and dynamic benchmarks overestimate LLM competence. Today's LLMs are still not trustworthy enough for real-world clinical deployment. To better meet clinical needs, future models should account for ambiguous inputs, patient misconceptions, and the messiness of real clinical data.

4.6 Case Study

Tabs.12–14 (in Appendix) compare the original case, versions generated by existing dynamic methods, and **DyReMe**'s version. DeepSeek-V3, GPT-4o, and Qwen3-32B diagnose the original and dynamic baseline versions correctly but misdiagnose **DyReMe**'s version. For instance (Tab.12), DeepSeek-V3 overemphasizes the symptom "*small pink rashes*," misdiagnosed as *contact dermatitis*.

Tabs.15–16 (in Appendix) show the evaluation results of GPT-4o (43.75) and WiNGPT2-9B (25.00). In this case, a patient self-diagnoses "*sideroblastic anemia*" and incorrectly assumes that "*Frequent sneezing is usually harmless and not a sign of a serious problem.*" GPT-4o provides the correct diagnosis (*iron deficiency anemia*), while WiNGPT2-9B incorrectly follows the patient's self-diagnosis. However, GPT-4o only partially corrects the misinformation, while WiNGPT2-9B fully addresses the health rumor. With regard to helpfulness, GPT-4o provides detailed evidence of the diagnosis and offers effective suggestions, such as "*avoid overexertion and maintain a good daily routine*", while WiNGPT2-9B only gives a brief suggestion of "*anemia treatment*". In terms of consistency, GPT-4o outputs three different diagnoses, while WiNGPT2-9B produces four, indicating greater consistency from GPT-4o. Overall, these case studies demonstrate that **DyReMe** generates more realistic and challenging questions, providing a clearer assessment of LLM trustworthiness.

5 Conclusion

Static medical diagnostic benchmarks often overestimate model ability and fail to reflect real-world complexity. To tackle this, we propose **DyReMe**—a dynamic framework for evaluating LLMs in medical diagnostics. **DyReMe** consists of two components: `DyGen`, which generates challenging and realistic questions by incorporating diagnostic distractors and varied expression styles, and `EvalMed`, which evaluates LLMs across *accuracy*, *veracity*, *helpfulness*, and *consistency*. Experiments show that **DyReMe** provides a more rigorous and realistic assessment compared to existing benchmarks. Results reveal that current LLMs still face limitations in realistic diagnostic scenarios, especially when dealing with misdiagnosis factors and lay-language expressions. These findings call for future work on improving LLM reliability and safety in real-world medical applications.

6 Limitations

We have only tested **DyReMe** on several public medical datasets and it cannot be directly applied to real-world clinical cases.

Multilingual Scenarios. Due to constraints in available datasets and the accessibility of diagnostic guidelines, our experiments are conducted exclusively in Chinese so far. Consequently, we have not yet examined the performance of **DyReMe** in other languages or in multilingual contexts. Nevertheless, given its high scalability, extending **DyReMe** into a dynamic evaluation framework that supports multilingualism is a promising direction for the future.

Self-Bias. The issue of self-bias—where an LLM evaluator might favor texts generated by itself—is still a matter of debate (Xie et al., 2024; Zheng et al., 2023). To address this concern, we have taken a cautious approach by: (1) using different LLMs for `DyGen` and `EvalMed`, (2) employing RAG to encourage the LLM to derive information from external references rather than relying solely on its internal knowledge, and (3) strictly scoring according to the defined rules. Results in Appendix E show that Self-Recognition (Davidson et al., 2024)'s impact is not significant.

Medical Multimodality. In real-world clinical practice, physicians rely on a variety of modalities—including medical images, laboratory results, and biosignals—for diagnosis. However, such multimodal data are limited and challenging to obtain. As a result, we only focus on text-based diagnostic scenarios in this work. With the rapid advancement of multimodal LLMs, future research could explore extending **DyReMe** to evaluate these models in multimodal medical contexts. Future work should aim to develop large-scale, diverse, and challenging evaluation benchmarks that incorporate these crucial multimodal inputs.

Scope and Clinical Validation. Our study focuses on a controlled consultation-style diagnostic setting to enable scalable and reproducible benchmarking. While **DyReMe** incorporates clinically grounded confounders, it does not fully capture end-to-end clinical workflows (e.g., longitudinal history, labs/imaging, follow-up, and multidisciplinary decision-making), nor does it constitute prospective clinical validation. Establishing real-world utility will require larger-scale studies with clinician-in-the-loop protocols and deployment-oriented evaluations in clinical settings.

7 Ethical Considerations

License. Our study utilizes four publicly available medical datasets, including RJUA (Lyu et al., 2024), DxBench (Chen et al., 2024a), DDXPlus (Tchango et al., 2022), and Dxy (Xu et al., 2019). All data is free of personally identifiable information, unique identifiers, and any offensive or objectionable content. The RJUA and DDXPlus data is published under a Creative Commons Attribution 4.0 International Licence (CC BY). DxBench and Dxy are distributed under the apache-2.0 license. We use data exclusively within the bounds of its license and solely for research purposes. In addition, we use the official open-source implementations provided by Self-Evolving (Wang et al., 2025) and DyVal2 (Zhu et al., 2024).

Potential Risks. Our system is designed exclusively for research and educational purposes. The medical advice provided is intended solely as a reference. Utilizing the system for clinical deployment carries inherent risks, as it has not undergone the rigorous validation required for direct application in patient care. Additionally, integrating LLMs with real medical data is highly sensitive. When testing with actual patient information, it is imperative to use HIPAA-compliant services—such as Azure OpenAI—to ensure robust protection of patient data and to adhere to privacy regulations.

References

Judith A Arter and Joseph R Jenkins. 1979. Differential diagnosis—prescriptive teaching: A critical appraisal. *Review of educational research*, 49(4):517–555.

Suhana Bedi and et al. 2025. MedHELM: Holistic Evaluation of Large Language Models for Medical Tasks. *Preprint*, arXiv:2505.23802.

Leah T Braun, Laura Zwaan, Jan Kiesewetter, Martin R Fischer, and Ralf Schmidmaier. 2017. Diagnostic errors by medical students: results of a prospective qualitative study. *BMC medical education*, 17(1):191.

Yan Cai, Linlin Wang, Ye Wang, Gerard de Melo, Ya Zhang, Yanfeng Wang, and Liang He. 2024. MedBench: A Large-Scale Chinese Benchmark for Evaluating Medical Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17709–17717.

Junying Chen, Chi Gui, Anningzhe Gao, Ke Ji, Xidong Wang, Xiang Wan, and Benyou Wang. 2024a. CoD, Towards an Interpretable Medical Agent using Chain of Diagnosis. *Preprint*, arXiv:2407.13301.

Junying Chen, Xidong Wang, Ke Ji, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, Haizhou Li, and Benyou Wang. 2024b. HuatuoGPT-II, One-stage Training for Medical Adaption of LLMs. *Preprint*, arXiv:2311.09774.

Qianyu Chen, Xin Li, Kunnan Geng, and Mingzhong Wang. 2023. Context-Aware Safe Medication Recommendations with Molecular Graph and DDI Graph Embedding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):7053–7060.

Philip Chung, Akshay Swaminathan, Alex J Goodell, Yeasul Kim, S Momsen Reincke, Lichy Han, Ben Deverett, Mohammad Amin Sadeghi, Abdel-Badih Ariss, Marc Ghanem, et al. 2025. Verifact: Verifying facts in llm-generated clinical text with electronic health records. *arXiv preprint arXiv:2501.16672*.

Tim R Davidson, Viacheslav Surkov, Veniamin Veselovsky, Giuseppe Russo, Robert West, and Caglar Gulcehre. 2024. Self-recognition in language models. *arXiv preprint arXiv:2407.06946*.

DingXiang. 2022. Tips for responding to five-star reviews. Accessed: 2025-07-26.

Alphaeus Dmonte, Roland Oruche, Marcos Zampieri, Prasad Calyam, and Isabelle Augenstein. 2024. Claim Verification in the Age of Large Language Models: A Survey. *Preprint*, arXiv:2408.14317.

Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. 2024. AI Hospital: Benchmarking Large Language Models in a Multi-agent Medical Interaction Simulator. *Preprint*, arXiv:2402.09742.

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

Tyler B. Forbush, Adi V. Gundlapalli, Miland N. Palmer, Shuying Shen, Brett R. South, Guy Divita, Marjorie Carter, Andrew Redd, Jorie M. Butler, and Matthew Samore. 2013. “Sitting on Pins and Needles”: Characterization of Symptom Descriptions in Clinical Notes”. *AMIA Summits on Translational Science Proceedings*, 2013:67–71.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling Synthetic Data Creation with 1,000,000,000 Personas. *Preprint*, arXiv:2406.20094.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations*.

Mark L Graber, Nancy Franklin, and Ruthanna Gordon. 2005. Diagnostic error in internal medicine. *Archives of internal medicine*, 165(13):1493–1499.

Vipul Gupta, Candace Ross, David Pantoja, Rebecca J. Passonneau, Megan Ung, and Adina Williams. 2024. Improving Model Evaluation using SMART Filtering of Benchmark Datasets. *Preprint*, arXiv:2410.20245.

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Elizabeth Healey, Amelia Li Min Tan, Kristen L. Flint, Jessica L. Ruiz, and Isaac Kohane. 2025. A case study on using a large language model to analyze continuous glucose monitoring data. *Scientific Reports*, 15(1):1143.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models. *Advances in Neural Information Processing Systems*, 36:62991–63010.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Minhao Jiang, Ken Ziyu Liu, Ming Zhong, Rylan Schaeffer, Siru Ouyang, Jiawei Han, and Sanmi Koyejo. 2024. Investigating Data Contamination for Pre-training Language Models. *Preprint*, arXiv:2401.06059.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Ariadna Juarez-Garcia, Tim Stokes, Beth Shaw, Janette Camosso-Stefinovic, and Richard Baker. 2006. The costs of epilepsy misdiagnosis in england and wales. *Seizure*, 15(8):598–605.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking Benchmarking in NLP. *Preprint*, arXiv:2104.14337.

Yunsoo Kim, Jinge Wu, Yusuf Abdulle, and Honghan Wu. 2024. MedExQA: Medical Question Answering Benchmark with Multiple Explanations. *Preprint*, arXiv:2406.06331.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Retno Larasati, Anna De Liddo, and Enrico Motta. 2023. Meaningful Explanation Effect on User's Trust in an AI Medical System: Designing Explanations for Non-Expert Users. *ACM Trans. Interact. Intell. Syst.*, 13(4):30:1–30:39.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Yucheng Li, Frank Guerin, and Chenghua Lin. 2024a. An Open Source Data Contamination Report for Large Language Models. *Preprint*, arXiv:2310.17589.

Yucheng Li, Yunhao Guo, Frank Guerin, and Chenghua Lin. 2024b. An Open-Source Data Contamination Report for Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 528–541, Miami, Florida, USA. Association for Computational Linguistics.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Jie Liu, Wenxuan Wang, Zizhan Ma, Guolin Huang, Yihang SU, Kao-Jung Chang, Wenting Chen, Haoliang Li, Linlin Shen, and Michael Lyu. 2024b. Medchain: Bridging the gap between llm agents and clinical practice through interactive sequential benchmarking. *arXiv preprint arXiv:2412.01605*.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, and Michael Lingzhi Li. 2023. Benchmarking Large Language Models on CMExam - A comprehensive Chinese Medical Exam Dataset. *Advances in Neural Information Processing Systems*, 36:52430–52452.

Siqi Luo, Hongyi Qin, Hanlin Li, and Cui Huang. 2024. Can ChatGPT provide health information as physicians do? Preliminary findings from a cross-sectional study of online medical consultation. *Information Research an international electronic journal*, 29(2):419–426.

Shiwei Lyu, Chenfei Chi, Hongbo Cai, Lei Shi, Xiaoyan Yang, Lei Liu, Xiang Chen, Deng Zhao, Zhiqiang Zhang, Xianguo Lyu, Ming Zhang, Fangzhou Li, Xiaowei Ma, Yue Shen, Jinjie Gu, Wei Xue, and Yiran Huang. 2024. RJUA-QA: A Comprehensive QA Dataset for Urology. *Preprint*, arXiv:2312.09785.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. *Preprint*, arXiv:2305.14251.

Nithesh Naik, B. M. Zeeshan Hameed, Dasharathraj K. Shetty, Dishant Swain, Milap Shah, Rahul Paul, Kaivalya Aggarwal, Sufyan Ibrahim, Vathsala Patil, Komal Smriti, Suyog Shetty, Bhavan Prasad Rai, Piotr Chlosta, and Bhaskar K. Somani. 2022. Legal and Ethical Consideration in Artificial Intelligence in Healthcare: Who Takes Responsibility? *Frontiers in Surgery*, 9.

Mathieu Nendaz and Arnaud Perrier. 2012. Diagnostic errors and flaws in clinical reasoning: Mechanisms and prevention in practice. *Swiss Medical Weekly*, 142(4344):w13706–w13706.

Robert Osazuwa Ness, Katie Matton, Hayden Helm, Sheng Zhang, Junaid Bajwa, Carey E. Priebe, and Eric Horvitz. 2024. MedFuzz: Exploring the Robustness of Large Language Models in Medical Question Answering. *Preprint*, arXiv:2406.06573.

Geoffrey R. Norman, Sandra D. Monteiro, Jonathan Sherbino, Jonathan S. Ilgen, Henk G. Schmidt, and Silvia Mamede. 2017. The Causes of Errors in Clinical Reasoning: Cognitive Biases, Knowledge Deficits, and Dual Process Thinking. *Academic Medicine*, 92(1):23.

OpenAI. 2025a. Gpt-5 system card. `https://cdn.openai.com/gpt-5-system-card.pdf`. Technical report.

OpenAI. 2025b. Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/. Accessed: 2025-07-27.

Daniele Orso, Nicola Federici, Roberto Copetti, Luigi Vetrugno, and Tiziana Bove. 2020. Infodemic and the spread of fake news in the covid-19-era. *European Journal of Emergency Medicine*, 27(5):327–328.

Zetian Ouyang, Yishuai Qiu, Linlin Wang, Gerard de Melo, Ya Zhang, Yanfeng Wang, and Liang He. 2024. CliMedBench: A Large-Scale Chinese Benchmark for Evaluating Medical Large Language Models in Clinical Scenarios. *Preprint*, arXiv:2410.03502.

Hao-Ting Pai, Wen-Cheng Chung, Xin-Hong Fang, Yu-Hsin Hsu, and Shu-Ting Huang. 2024. The Explainable Analytics for Exploring Misdiagnoses. In *Proceedings of the 2024 8th International Conference on Medical and Health Informatics*, ICMHI '24, pages 238–243, New York, NY, USA. Association for Computing Machinery.

Jiazhen Pan, Bailiang Jian, Paul Hager, Yundi Zhang, Che Liu, Friedrike Jungmann, Hongwei Bran Li, Chenyu You, Junde Wu, Jiayuan Zhu, et al. 2025. Beyond benchmarks: Dynamic, automatic and systematic red-teaming agents for trustworthy medical language models. *arXiv preprint arXiv:2508.00923*.

Ye-Jean Park, Abhinav Pillai, Jiawen Deng, Eddie Guo, Mehul Gupta, Mike Paget, and Christopher Naugler. 2024. Assessing the research landscape and clinical utility of large language models: a scoping review. *BMC Medical Informatics and Decision Making*, 24(1):72.

Leili Pourafkari, Arezou Tajlil, Samad Ghaffari, Rezayat Parvizi, Mohammadreza Chavoshi, Kasra Kolahdouzan, Nasrin Khaki, Raziyeh Parizad, Geoffery G. Hobika, and Nader D. Nader. 2017. The frequency of initial misdiagnosis of acute aortic dissection in the emergency department and its impact on outcome. *Internal and Emergency Medicine*, 12(8):1185–1195.

Jiaxing Qiu, Dongliang Guo, Papini Natalie, Peace Noelle, Levinson Cheri, and Teague R. Henry. 2025. Ensemble of Large Language Models for Curated Labeling and Rating of Free-text Data. *Preprint*, arXiv:2501.08413.

Winning Health AI Research. 2025. Wingpt2. Accessed: 2025-07-26.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

C. E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Tian Shen, Yu Li, and Xi Chen. 2024. A Systematic Review of Online Medical Consultation Research. *Healthcare*, 12(17):1687.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. VERISCORE: Evaluating the factuality of verifiable claims in long-form text generation. *Preprint*, arXiv:2406.19276.

David M. Studdert, Michelle M. Mello, William M. Sage, Catherine M. DesRoches, Jordon Peugh, Kinga Zapert, and Troyen A. Brennan. 2005. Defensive Medicine Among High-Risk Specialist Physicians in a Volatile Malpractice Environment. *JAMA*, 293(21):2609–2617.

Xiaoyu Tan, Shaojie Shi, Xihe Qiu, Chao Qu, Zhenting Qi, Yinghui Xu, and Yuan Qi. 2023. Self-criticism: Aligning large language models with their understanding of helpfulness, honesty, and harmlessness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 650–662, Singapore. Association for Computational Linguistics.

Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. 2022. DDXPlus: A New Dataset For Automatic Medical Diagnosis. *Preprint*, arXiv:2205.09148.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

thisisjeffchen. 2023. ddxplus-parser/ddxplus_parser.py at main · thisisjeffchen/ddxplus-parser. [Online; accessed 2025-07-30].

Ehsan Ullah, Anil Parwani, Mirza Mansoor Baig, and Rajendra Singh. 2024. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology – a recent scoping review. *Diagnostic Pathology*, 19(1):43.

VolcEngine. 2025a. doubao-pro-32k system card. [Online; accessed 2025-07-31].

VolcEngine. 2025b. Functionality overview of the content connectivity plugin. [Online; accessed 2025-07-27].

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Graham Neubig, Yonatan Bisk, and Hao Zhu. 2024a. SOTOPIA-π: Interactive Learning of Socially Intelligent Language Agents. https://arxiv.org/abs/2403.08715v3.

Siyuan Wang, Zhuohan Long, Zhihao Fan, Xuanjing Huang, and Zhongyu Wei. 2025. Benchmark self-evolving: A multi-agent framework for dynamic LLM evaluation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3310–3328, Abu Dhabi, UAE. Association for Computational Linguistics.

Xidong Wang, Guiming Chen, Song Dingjie, Zhang Zhiyi, Zhihong Chen, Qingying Xiao, Junying Chen, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2024b. CMB: A Comprehensive Medical Benchmark in Chinese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6184–6205, Mexico City, Mexico. Association for Computational Linguistics.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. *Preprint*, arXiv:2411.04368.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. 2024. LiveBench: A Challenging, Contamination-Free LLM Benchmark. https://arxiv.org/abs/2406.19314v1.

Jiaxin Wu, Yizhou Yu, and Hong-Yu Zhou. 2024. Uncertainty Estimation of Large Language Models in Medical Question Answering. *Preprint*, arXiv:2407.08662.

Rui Wu, Zhaopeng Qiu, Jiacheng Jiang, Guilin Qi, and Xian Wu. 2022. Conditional Generation Net for Medication Recommendation. In *Proceedings of the ACM Web Conference 2022*, pages 935–945, Virtual Event, Lyon France. ACM.

Yiqing Xie, Sheng Zhang, Hao Cheng, Pengfei Liu, Zelalem Gero, Cliff Wong, Tristan Naumann, Hoifung Poon, and Carolyn Rose. 2024. DocLens: Multi-aspect Fine-grained Medical Text Evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 649–679, Bangkok, Thailand. Association for Computational Linguistics.

Cheng Xu, Shuhao Guan, Derek Greene, and M.-Tahar Kechadi. 2024. Benchmark Data Contamination of Large Language Models: A Survey. *Preprint*, arXiv:2406.04244.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-End Knowledge-Routed Relational Dialogue System for Automatic Diagnosis. *Preprint*, arXiv:1901.10623.

Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. 2024. To Believe or Not to Believe Your LLM. *Preprint*, arXiv:2406.02543.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Chaoqi Yang, Cao Xiao, Fenglong Ma, Lucas Glass, and Jimeng Sun. 2021. SafeDrug: Dual Molecular Graph Encoders for Recommending Effective and Safe Drug Combinations. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3735–3741, Montreal, Canada. International Joint Conferences on Artificial Intelligence Organization.

Yefei Yang, Xiaofei Zhang, and Peter K. C. Lee. 2019. Improving the effectiveness of online healthcare platforms: An empirical study with multi-period patient-doctor consultation data. *International Journal of Production Economics*, 207:70–80.

John Zarocostas. 2020. How to fight an infodemic. *The lancet*, 395(10225):676.

Qing T. Zeng and Tony Tse. 2006. Exploring and developing consumer health vocabularies. *Journal of the American Medical Informatics Association: JAMIA*, 13(1):24–29.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, et al. 2023. Huatuogpt, towards taming language model to be a doctor. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 10859–10885.

Ming Zhang, Yujiong Shen, Zelin Li, Huayu Sha, Binze Hu, Yuhui Wang, Chenhao Huang, Shichun Liu, Jingqi Tong, Changhao Jiang, Mingxu Chai, Zhiheng Xi, Shihan Dou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025. LLMEval-Med: A Real-world Clinical Benchmark for Medical LLMs with Physician Validation. *Preprint*, arXiv:2506.04078.

Zhehao Zhang, Jiaao Chen, and Diyi Yang. 2024. Darg: Dynamic evaluation of large language models via adaptive reasoning graph. *Advances in Neural Information Processing Systems*, 37:135904–135942.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Preprint*, arXiv:2306.05685.

Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. 2023. DyVal: Dynamic Evaluation of Large Language Models for Reasoning Tasks. In *The Twelfth International Conference on Learning Representations*.

Kaijie Zhu, Jindong Wang, Qinlin Zhao, Ruochen Xu, and Xing Xie. 2024. Dynamic Evaluation of Large Language Models by Meta Probing Agents. In *Forty-First International Conference on Machine Learning*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A Benchmarking Platform for Text Generation Models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 1097–1100, New York, NY, USA. Association for Computing Machinery.

A Traps Description

Below is an introduction to four types of traps:

- **Self-Diagnosis**: Introduce patient-suggested diagnoses to simulate confirmation bias and test the LLM's ability to consider alternative possibilities.
- **Distracting History**: Include irrelevant medical history to obscure key information and mimic over-interpretation bias.
- **External Noise**: Add non-critical external factors, such as environmental or lifestyle details, to replicate background noise bias.
- **Symptom Misplaced**: Blur the distinction between primary and secondary symptoms, challenging the model's ability to focus on critical diagnostic clues.

B Experiments Setup

B.1 Dataset and Hyperparameter Details

We have chosen three diagnostic datasets:

- **DDXPlus** (Tchango et al., 2022) offers a synthetic dataset of 1.3 million patients, featuring differential diagnoses along with real patient pathology, symptoms, and medical history. We use the English version and translate it into Chinese to ensure consistency with other datasets.
- **DxBench** (Chen et al., 2024a) is a real-world Chinese diagnostic benchmark comprising 1,148 actual cases spanning 461 diseases.
- **Dxy** (Xu et al., 2019) is a Chinese diagnostic dataset sourced from an online health community, comprising 104 samples. We utilize the version provided by (Chen et al., 2024a).

These three datasets are selected for their diverse sources and characteristics, ensuring a comprehensive evaluation of LLMs in medical diagnostics. For all datasets, we use their development and test sets. To align with the LLM diagnostic pattern, we transform the original symptom data into a patient inquiry format (Fig.7). To streamline development, we filtered out samples with multiple true diagnoses. DxBench contains 1148 questions, while Dxy has 104 questions. For DDXPlus, we randomly select 300 questions from the original dataset to ensure a manageable size for development.

Across all experiments, we consistently set the generation temperature at 0.7, use a temperature of 0 for question verification, and maintain a temperature of 0 during evaluation. For DyVal2 (Zhu et al., 2024) and Self-Evolving (Wang et al., 2025), we use their official implementation. Since they were designed for multiple-choice questions rather than Q&A tasks, we remove processes that do not apply to the diagnostic Q&A task during experiments. To benchmark 12 leading LLMs, we generate 3,200 questions based on 800 seed questions from DxBench, using Acc($\cdot$), Ver($\cdot$), Help($\cdot$), and Cons($\cdot$) as metrics. Alg.1 and Alg.2 detail the DyGen and EvalMed algorithms, respectively.

Algorithm 1 DyGen: Dynamic Question Generation

Input seed dataset $D = \{(q_i,\, d_i,\, m_i)\}$ ▷ q: question, d: true diagnosis, m: medical entity (random symptom)
Input trap pool S, persona pool B
Input validators V, refiner R
Input helpfulness criteria set H
Output benchmark dataset $\mathcal{Q}$

$\mathcal{Q} \leftarrow \varnothing$
for all $(q, d, m) \in D$ **do**
  $d_{\text{dis}} \leftarrow \mathcal{G}_{\text{RAG}}(d)$
  $q_{\text{trap}} \leftarrow \textsc{ApplyTrap}(q,\, S,\, d_{\text{dis}})$
  $q_{\text{per}} \leftarrow \textsc{ApplyStyle}(q_{\text{trap}},\, B)$
  $(e_{rumor}, e_{fact}) \leftarrow \textsc{GenRumorSnippet}(d,\, m)$
  $q_{\text{rum}} \leftarrow \textsc{InsertRumor}(q_{\text{per}},\, e_{rumor})$
  $q^{\star} \leftarrow \textsc{ValidateAndRefine}(q_{\text{rum}},\, V,\, R)$
  $K \leftarrow \varnothing$
  for all $h \in H$ **do**
    $k_h \leftarrow \textsc{DeriveKey}(d,\, q^{\star},\, h)$
    $K[h] \leftarrow k_h$
  end for
  $\mathcal{Q}.\textsc{append}\big(q^{\star},\, d,\, (e_{rumor}, e_{fact}),\, K\big)$
end for
return $\mathcal{Q}$

B.2 Challenge Experiment Setup

During question generation, both for **DyReMe** and the current dynamic evaluation methods, we employ GPT-4.1 (OpenAI, 2025b) as the generator, with a generation temperature of 0.7 and a verification temperature of 0. For challenge evaluation, we adopt GPT-4.1 as the judge, setting its tem-

Algorithm 2 EvalMed: Comprehensive Model Evaluation

```
Input  model M, benchmark dataset Q helpfulness criteria set H, weights w
Output  scores (Acc, Ver, Help, Cons)
Acc ← 0, Ver ← 0, Help ← 0
P ← ∅
for all (q, d, r, K) ∈ Q do
    â ← M(q)
    Acc += Φ_acc(â, d)
    Ver += Φ_ver(â, r, d)
    for all h ∈ H do
        Help += w_h Φ_h(â, K[h])
    end for
    P.append(â)
end for
Cons ← ConsistencyMetric(P)
return (Acc/|Q|, Ver/|Q|, Help/(|Q| · |H|), Cons)
```

perature to 0. We select 11 baselines. For every baseline LLM, we set the response temperature to 0 to ensure optimal performance. We measure challenge use the average of Top-1,3,5 diagnostic accuracy. Specifically, we instruct the LLM to provide a ranked list of diagnoses, and we judge whether the true diagnosis appears in the top 1, 3, or 5 positions. To mitigate randomness, we adopt the boost-trap method for hypothesis testing, following precedents in medical AI research (Yang et al., 2021; Wu et al., 2022; Chen et al., 2023). Specifically, in each iteration, we randomly sample 80% of the evaluation dataset to form a subset, calculate the LLM's response accuracy on that subset, and repeat the process 10 times. We then conduct a one-sided t-test on these 10 sets of results to evaluate **DyReMe**'s challenge advantage over the state-of-the-art method. Figs.17-18 are prompts used in challenge assessment. The p-values are reported in Tab.7.

B.3 Diversity Experiment Setup

In our experiments, we evaluate two types of diversity: **expression diversity** and **diagnosis diversity**. The procedures are detailed below.

Expression Diversity

1. **Extraction:** For each question, use the an LLM to extract the style features. The features have three dimensions (each dimension has three level): *medical knowledge* (low, medium, high), *clarity* (low, medium, high), and *communication style* (indirect, neutral, direct).

2. **Computing Entropy:** For each type, we compute the entropy of the level's distribution:

$$H(X) = -\sum_{i=1}^{n} p_i \log_2(p_i),$$

where p_i is the proportion of the i-th level in the distribution. Then we compute the average entropy across all three types:

$$D_{\text{exp}} = \frac{1}{3}\sum_{i=1}^{3} H(X_i),$$

where X_i is the distribution of the i-th type.

The following pseudocode formalizes the above process:

Algorithm 3 Compute Expression Diversity

```
Input  Q                                   ▷ set of questions
Initialize counts cnt_M, cnt_C, cnt_S to zero
for q ∈ Q do
    (m, c, s) ← ExtractStyle(q)
    cnt_M[m]++, cnt_C[c]++, cnt_S[s]++
end for
function ENTROPY(cnt, N)
    H ← 0
    for each level ℓ with p = cnt[ℓ]/N > 0 do
        H −= p log_2 p
    end for
    return H
end function
N ← |Q|
D_exp ← (Entropy(cnt_M, N) + Entropy(cnt_C, N) + Entropy(cnt_S, N))/3
return D_exp
```

In the process of computing diversity, we utilize GPT-4.1 (OpenAI, 2025b) as the worker LLM. Figs.19-20 are prompts used in diversity assessment.

B.4 Benchmark Experiment Setup

We randomly select 800 questions from DxBench (Chen et al., 2024a) and generate 4 new questions for each using four types of traps (Appendix A), resulting in a benchmark with 3200 new questions. Our observations are conducted on 12 baseline LLMs: GPT-4o-2024-11-20 (Hurst et al., 2024), GPT-4o-mini-2024-07-18 (Hurst et al., 2024), o1-2024-12-47 (Jaech et al., 2024), o1-mini-2024-09-12 (Jaech et al., 2024), DeepSeek-V3-0324 (Liu et al., 2024a), Qwen3-32B (Team, 2025), Qwen2.5-32B-Instruct (Yang et al., 2024), Qwen2.5-7B-Instruct (Yang et al., 2024), WiNGPT2-Gemma-2-9B-Chat (Research, 2025), GLM-4-32B-0414 (GLM et al., 2024), Gemma-3-27B-it (Team et al., 2025), and MedGemma-27B-text-it (Sellergren et al., 2025). Figs.7-16 are prompts used for question generation. Figs.21-23 are prompts used in benchmark assessment. For all evaluation metrics, we use GPT-4.1 (OpenAI, 2025b) as the judge model, with a temperature of 0 for all evaluations. The max times of optimization iteration is set to 3. The number of scorepoints is set to 3. When answering questions, the number of max tokens is set to 2048, the number of sampling is 1, and in the prompt, we instruct the LLM not to output more than 200 words to ensure concise responses. We use personas from PersonaHub (Ge et al., 2024). For accuracy, we use the Top-1 accuracy as the final score. For veracity, we use the proportion of LLM rectifying rumors as the final veracity score.

For veracity, we generate up to 10 rumor-fact pairs for a medical entity (e.g., a symptom). We inject a rumor into each case to create a controlled stress-test setting, rather than implying that rumors occur in every real-world consultation. To minimize additional anchoring bias, each rumor is constructed strictly from the patient's reported symptoms (without introducing extra diagnostic cues beyond the case itself).

For helpfulness, with each item rated on 0–100, we use the average score of h_{evi}, h_{treat}, and h_{life} as the final helpfulness score. We control the number of score-points at 3.

For consistency, we conduct normalization to map the entropy value to a 0–100 scale, where 0 indicates maximum inconsistency and 100 indicates maximum consistency. Diagnoses are obtained via an LLM-based extractor (Fig. 20) and then standardized by a second LLM pass (Fig. 23) to merge surface variants (e.g., GERD vs. gastroesophageal reflux disease); for multi-diagnosis answers, we retain a fixed top diagnosis per response, consistent with our single-diagnosis benchmark.

Detailed standard deviation is shown in Tab.2.

B.5 Human Study Experiment Setup

We invited three clinical doctoral students majoring in medicine from a QS-ranked Top 100 medical college to act as medical annotators for expertise and reliability. All of them have at least one year of clinical experience. Given that analyzing a medical question and two responses is time-consuming (around 2 minutes per instance), it is challenging to further scale up the number of questions. Tab.11 shows the failure case of differential diagnoses retrieval. Most cases are caused by the difference in scale type or lesion sites.

We also randomly sampled 30 sets of scoring points covering different categories and evaluated their consistency with the medical encyclopedia (four levels: inconsistent, low, medium, and high). Of these, 25 sets are highly consistent, 4 are moderately consistent, and 1 shows low consistency. The lone low-consistency case is the lifestyle advice for *roseola infantum*: "maintain good personal hygiene, wash hands frequently (seven-step method), avoid scratching the rash, and cleanse the skin with warm water, keeping it dry", which is not explicitly covered by the encyclopedia.

Before the annotation tasks, we briefed the experts on the diagnostic scope of the questions and ensured they possessed the necessary diagnostic knowledge. We provided a detailed explanation of our research objectives, experimental setup, and data usage, and explicitly obtained informed consent from all experts. A pilot annotation showed that the average annotation time per expert for both tasks was about 60 minutes, for which we provided a remuneration of 100 RMB ($14) per expert. Data contains no protected information. Fig.5 display the screenshot of the annotation platform and Fig.5 is its English translation. We use Gwet's AC1 (Gwet, 2008) as our agreement metric, as it is widely adopted (Chung et al., 2025; Qiu et al., 2025; Healey et al., 2025) and mitigates the adverse effects of imbalanced distributions. Results of all annotation tasks are detailed in Tab.3-5.

C Computational Experiments Details

Model Size And Budget. All open-source models in our paper are executed (with vLLM (Kwon et al., 2023)) on a server with inference for the benchmark experiment finished in roughly 1 day. Tab.6 details the parameters of the open-source LLMs. For closed-source LLMs, we get responses via API calls.

Model	acc	ver	help	cons	avg
Qwen3-32B	0.3162	0.3860	0.1200	0.4685	0.3227
GPT-4o	0.4223	0.4506	0.1668	0.6934	0.4333
DeepSeek-V3	0.4180	0.4567	0.1324	0.2711	0.3195
MedGemma-27B	0.4287	0.3819	0.1479	0.6227	0.3953
Gemma-3-27B	0.5438	0.4279	0.1767	0.5622	0.4277
o1-mini	0.3583	0.5893	0.1341	0.5857	0.4168
o1	0.4384	0.3196	0.1605	0.3577	0.3190
Qwen2.5-32B	0.4372	0.3077	0.1267	0.6879	0.3899
GPT-4o-mini	0.3112	0.4951	0.1440	0.4143	0.3411
Qwen2.5-7B	0.3460	0.4171	0.2004	0.2519	0.3038
GLM-4-32B	0.5426	0.4910	0.1219	0.2534	0.3522
WiNGPT2-9B	0.3604	0.4231	0.1419	0.5242	0.3624

Table 2: Standard deviation of benchmark results for various models

ID	Expert-1	Expert-2	Expert-3
Q_1	4	4	3
Q_2	4	5	4
Q_3	5	4	4
Q_4	4	4	4
Q_5	3	4	4
Q_6	5	3	5
Q_7	2	3	3
Q_8	4	5	4
Q_9	4	4	4
Q_{10}	4	3	4
Q_{11}	3	4	2
Q_{12}	5	5	4
Q_{13}	4	3	4
Q_{14}	4	5	4
Q_{15}	5	5	2
Q_{16}	4	4	4
Q_{17}	3	3	2
Q_{18}	5	2	4
Q_{19}	4	4	5
Q_{20}	5	4	4
Q_{21}	3	4	4
Q_{22}	5	5	5
Q_{23}	4	5	4
Q_{24}	4	4	4
Q_{25}	4	4	5
Q_{26}	5	4	4
Q_{27}	4	4	4
Q_{28}	4	5	2
Q_{29}	4	4	4
Q_{30}	4	4	4

Table 3: Expert ratings on the *Question Quality Task* for the original (pre-rewrite) questions, used as a scoring baseline. These questions are drawn from a disjoint set and do not correspond one-to-one with those in Tabs. 4 (no case overlap).

ID	Expert-1	Expert-2	Expert-3
Q_1	5	4	5
Q_2	4	4	5
Q_3	5	4	3
Q_4	4	4	2
Q_5	4	4	2
Q_6	5	3	2
Q_7	5	4	2
Q_8	5	4	4
Q_9	4	4	3
Q_{10}	5	2	2
Q_{11}	5	2	2
Q_{12}	5	3	2
Q_{13}	5	4	3
Q_{14}	4	4	3
Q_{15}	5	4	2
Q_{16}	4	5	5
Q_{17}	4	4	3
Q_{18}	4	4	3
Q_{19}	5	4	3
Q_{20}	5	4	3
Q_{21}	5	4	4
Q_{22}	5	2	4
Q_{23}	5	4	3
Q_{24}	5	4	4
Q_{25}	5	4	4
Q_{26}	5	4	3
Q_{27}	5	4	3
Q_{28}	5	4	4
Q_{29}	5	4	4
Q_{30}	5	4	3

Table 4: Experts' results on *Question Quality Task* (30 questions)

ID	Expert-1	Expert-2	Expert-3
Q_1	1	1	1
Q_2	1	1	1
Q_3	1	1	1
Q_4	0	1	1
Q_5	1	1	1
Q_6	1	1	1
Q_7	1	1	1
Q_8	1	0	1
Q_9	1	1	1
Q_{10}	1	1	0
Q_{11}	1	0	1
Q_{12}	1	1	1
Q_{13}	1	1	1
Q_{14}	1	1	1
Q_{15}	1	1	1
Q_{16}	1	1	1
Q_{17}	1	1	1
Q_{18}	1	1	1
Q_{19}	1	1	1
Q_{20}	1	1	1
Q_{21}	1	1	1
Q_{22}	1	0	1
Q_{23}	1	1	1
Q_{24}	1	1	1
Q_{25}	1	1	1
Q_{26}	1	1	1
Q_{27}	1	1	1
Q_{28}	1	1	1
Q_{29}	1	1	1
Q_{30}	1	1	1

Table 5: Agreement between experts and Dyreme outputs (1 = agreement, 0 = disagreement).

Descriptive Statistics. For the challenging, we utilize the boosttrap method described in Appendix B.2 to conduct 10 sampling runs. Tab.7 reports the p-values, verified by a one-tailed t-test. Each column shows the p-value for using the best baseline method as the control group with **DyReMe** as the experimental group. For diversity, extensibility, and ablation study, due to some metrics (e.g., expression diversity, diagnosis diversity) need to be computed on the entire dataset, we only report the final results. For the benchmark experiment, we report the average performance and the standard deviation.

Parameters For Packages. We adopt the Self-BLEU (Zhu et al., 2018) approach, employing NLTK's `sentence_bleu` function for Self-BLEU computation with identical parameters. We utilize ddxplus_parser(thisisjeffchen, 2023) to process the original DDXPlus (Tchango et al., 2022) data. In addition, we use the official open-source implementations provided by Self-Evolving (Wang et al., 2025) and DyVal2 (Zhu et al., 2024).

Model	**Size**
Qwen3-32B	32B
Qwen2.5-32B-Instruct	32B
Qwen2.5-7B-Instruct	7B
Gemma-3-27B-it	27B
MedGemma-27B-text-it	27B
WiNGPT2-Gemma-2-9B	9B
GLM-4-32B-0414	32B

Table 6: Open-source LLM size.

Model	**DyReMevs. Runner-up** (p-value)
DeepSeek-V3	<0.001
GPT-4o	<0.001
GPT-4o-mini	<0.001
MedGemma-27B	<0.001
WiNGPT2-9B	<0.001
Qwen3-32B	<0.001
Gemma-3-27B	<0.001
GLM-4-32B	<0.001
Qwen2.5-32B	<0.001
Qwen2.5-7B	<0.001

Table 7: Statistical significance (p-value) between the **DyReMe** (experimental group, **bold** in Table 1) and the strongest baseline (runner-up, underlined) for each model. All results are statistically significant with $p <$ 0.001.

D Generative Assistance

We employ AI assistants to refine our paper's content, and we leverage AI tools to aid in code development.

E Reliability Checks

We perform two reliability checks: (i) generator-side self-recognition, and (ii) evaluator robustness via a judge-swap stress test.

Generator-side self-recognition. We conduct experiments on **DxBench** (Chen et al., 2024a) and **RJUA** (Lyu et al., 2024). DxBench covers 461 diseases with 1,148 questions, and RJUA is a urology QA dataset curated by domain experts with 344 questions. For each dataset, we generate two benchmarks using `DouBao-Pro-32K-241215` (VolcEngine, 2025a) and `DeepSeek-V3-241226` (Liu et al., 2024a), respectively, and compare Top-1 diagnostic accuracy across the two benchmarks (Table 8). If self-recognition is present, a model should show a systematic advantage on the benchmark it generated (i.e., a diagonal boost). Instead, we observe no such pattern: DouBao does not perform better on DxBench-DouBao than on DxBench-DeepSeek (22.66 vs. 25.98), and the performance gap between DouBao and DeepSeek remains similar across benchmarks (e.g., 3.22 on DxBench-DeepSeek vs. 3.79 on DxBench-DouBao). Overall, these results suggest that generator-side self-recognition is not significant in our setting.

	DxBench		RJUA	
Model	DeepSeek	DouBao	DeepSeek	DouBao
DouBao	25.98	22.66	36.62	38.07
DeepSeek	29.20	26.45	36.65	37.23

Table 8: Generator-side self-recognition analysis (Top-1 diagnostic accuracy).

Evaluator robustness (judge-swap stress test). To test whether our findings depend on a specific evaluator, we swap the judge from GPT-4.1 to GPT-5 and re-run the challenge evaluation (Tables 9–10). While absolute scores change under a different judge, the relative difficulty is preserved: DyReMe consistently yields the largest performance drops (about 11–12%) across all models, whereas Self-Evolving induces only minor changes (about 0–1%). The preservation of these trends under a stronger, independent judge indicates that our conclusions are not tied to a particular evaluator and instead reflect the intrinsic difficulty of the generated questions.

Model	DDXPlus	DxBench	Dxy	Avg.	Self-Evolv. Δ	DyVal2 Δ	DyReMe Δ
DeepSeek-V3	80.78	70.50	77.02	72.92	73.13 (+0.29%)	69.50 (-4.69%)	65.26 (-10.51%)
GPT-4o	81.11	70.15	74.11	72.53	72.98 (+0.62%)	69.67 (-3.94%)	64.74 (-10.75%)
Qwen2.5-32B	70.56	66.61	74.76	67.92	69.35 (+2.11%)	66.30 (-2.38%)	60.05 (-11.59%)

Table 9: Challenge results with GPT-4.1 as the judge (Tab. 1 in the main paper).

Model	DDXPlus	DxBench	Dxy	Avg.	Self-Evolv. Δ	DyVal2 Δ	DyReMe Δ
DeepSeek-V3	75.33	63.65	76.05	71.68	67.03 (+0.44%)	63.90 (-4.26%)	59.11 (-11.68%)
GPT-4o	75.11	63.50	73.46	70.69	67.10 (+1.03%)	64.48 (-2.92%)	58.53 (-11.97%)
Qwen2.5-32B	64.22	59.47	74.43	66.04	62.02 (+1.04%)	59.26 (-3.47%)	54.47 (-11.25%)

Table 10: Challenge results with GPT-5 as the judge (judge-swap stress test).

#	d_{org}	d_{dis}	Failure reason
1	Seborrheic dermatitis	Scalp/facial psoriasis	Scale type and lesion borders differ greatly; little overlap.
2	Tic disorder	Wilson's disease	Symptom profiles are clearly distinct.
3	Hypothyroidism	Sleep-apnea syndrome	Symptom profiles are clearly distinct.
4	Herpes zoster (shingles)	Herpes simplex	Lesion sites differ; minimal overlap expected.

Table 11: Examples of failed cases


医疗诊断问题评估问卷

本问卷包含两个任务:

任务一：问题质量评分 (1-5分)

请评估问题的质量（包含临床合理性和问题的临床代表性）：
- 1分：完全不合理，与真实临床场景不符
- 3分：合理性和代表性中等，与真实临床场景部分符合
- 5分：非常合理，与真实临床场景高度符合

任务二：选择更可信的回复

请从两个回复中选出更可信的一项。可信的含义包括:
1. 是否存在事实性错误?
2. 是否给出正确的诊断?
3. 诊断依据是否合理?
4. 是否包含合理的治疗和生活建议?

警告：问卷中可能包含让人不适的内容。

在完成问卷时,

注意：您可以参阅专业书籍、指南和来自互联网的可信信息源辅助判断

微信昵称

请输入微信昵称

问题 1
题目描述： 最近我经常出现腮腺肿胀和疼痛，感觉像是发炎，有时腮腺管也像被堵住一样，每次发作后隔一段时间又会复发。体检时还发现白细胞有点偏低。另外，我父亲以前得过结核，但我自己没有咳嗽或咳痰的症状。我想问，这种反复腮腺肿胀和疼痛会不会和家里有结核史有关？白细胞偏低和这些症状有没有关系？我在《国际炎症医学杂志》上看到说炎症症状如果一周内没有加重一般可以自愈，那我是不是可以先自己观察，不用急着去医院？这种情况平时应该怎么处理和预防?
真实诊断： 慢性复发性腮腺炎

Figure 5: The screenshot of the human annotation platform.

Original:
I have recently been experiencing a number of physical discomforts at the same time. First, I have noticeable sharp pain in both of my breasts; the pain level is approximately moderate, but it does not radiate to other areas. The pain did not occur particularly suddenly, and its location is not very precise. In addition, I have developed rapid breathing and difficulty breathing, accompanied by coughing, and the color of the sputum I cough up has changed and its volume has increased. Sometimes I also feel chills and shiver. My skin has developed pink rashes, with some areas showing swelling and peeling; the rashes are located on the back of my neck, behind my right ear, on the right side of my neck, on my left buttock, and on my left waist, but each rash is small, with a diameter of less than one centimeter. The rashes themselves are not very painful, only mildly uncomfortable, but the itching is quite pronounced, at a moderate level. Recently I have also often felt muscle aches all over my body, felt extremely fatigued, generally unwell, and my appetite has decreased significantly. It should be noted that I have a history of Parkinson's disease, chronic obstructive pulmonary disease, asthma, and heart failure; I have used bronchodilators, have had pneumonia, and underwent surgery in the past month. In addition, my weight has always been on the low side, with a BMI below 18.5. I have not traveled abroad in the last four weeks. May I ask what might be causing these symptoms? What examinations are needed or how should they be treated? Do I need to seek medical attention urgently?

Ground-Truth Diagnosis: *Pneumonia*

DeepSeek's Diagnosis: *Pneumonia* 😊

Self-Evolving:
At present, I have moderate sharp pain in both breasts, the location is relatively clear, the onset is gradual, and it does not radiate to other areas. At the same time, I experience rapid and difficult breathing, persistent coughing, and I am able to cough up colored or increased amounts of sputum; I also occasionally shiver. I have been diagnosed with Parkinson's disease and heart failure, have had pneumonia, and have a history of chronic obstructive pulmonary disease and asthma, and have used bronchodilators. Recently my skin has shown issues: pink rashes on the back of my neck, behind my right ear, on the right side of my neck, on my left buttock, and on my left waist; the rashes may peel, are moderately swollen, mildly painful, each lesion is under one centimeter in diameter, and they itch intensely at a moderate level. Additionally, I have whole-body muscle pain, poor appetite, a sharp reduction in food intake, extreme fatigue, general discomfort, and muscle soreness. I had surgery just last month and have had an ongoing cough problem. I have not traveled abroad in the past four weeks, and my body weight is low. May I ask what is causing these symptoms? Are they related to each other? How should they be treated and alleviated?

Ground-Truth Diagnosis: *Pneumonia*

DeepSeek's Diagnosis: *Pneumonia* 😊

DyVal2:
Normally I like to go for walks in the park on weekends, usually walking for about an hour each time. Recently I also tried learning a new language, spending half an hour each day practicing. Within this daily activity context, I am currently feeling sharp pain in both breasts, at a moderate level, not radiating to other areas, with a relatively clear location and gradual onset. Meanwhile, I have rapid and difficult breathing, continuous coughing, and am able to cough up unusually colored or increased quantities of sputum, and occasionally shiver. I was previously diagnosed with Parkinson's disease and heart failure, have had pneumonia, suffer from chronic obstructive pulmonary disease and asthma, and have used bronchodilators. Recently my skin has shown issues: small pink rashes on the back of my neck, behind my right ear, on the right side of my buttock, and on my left waist; the rashes may peel, are moderately swollen, lightly painful, each lesion is under one centimeter in diameter, and they are extremely itchy at a moderate level. In addition, I have whole-body muscle pain, no appetite, a large reduction in food intake, extreme fatigue, general discomfort, and muscle soreness. I had surgery last month and have been coughing all along. I have not traveled abroad in the past four weeks, and my weight is low (BMI under 18.5), with a recent slight decrease. What is the cause of these symptoms? How are they related? How should they be treated and alleviated?

Ground-Truth Diagnosis: *Pneumonia*

DeepSeek's Diagnosis: *Pneumonia* 😊

DyReMe:
Recently I have noticed some small pink rashes, each under one centimeter in diameter, on the back of my neck, behind my right ear, on the right side of my neck, on my left buttock, and on my left waist, accompanied by mild swelling, peeling, and significant itching, though the rashes themselves are not painful. I'm a bit worried if there is any skin problem. At the same time, I have been experiencing whole-body muscle aches, feeling particularly fatigued, and my appetite has also declined a lot, so I have been in a poor state overall. Occasionally I also feel my breathing is not very smooth, and I cough and produce phlegm, but without particularly severe breathing difficulty. I have a history of Parkinson's disease, heart failure, COPD, and asthma, and recently had surgery; my weight has always been low (BMI below 18.5), and it has declined a bit further lately. However, I understand that I don't need to worry too much about the weight change. I want to ask: what might be the cause of these rashes and systemic symptoms? What examinations are needed? Should I go to the hospital immediately given my current condition?

Ground-Truth Diagnosis: *Pneumonia*

DeepSeek's Diagnosis: *Contact dermatitis* 😳

Table 12: A Case (*Pneumonia*) from DDXPlus. To facilitate reading, we translate the questions from Chinese into English.

Original:
Recently, my child has not been feeling well, mainly with some discomfort in the throat, always feeling a bit blocked or painful. There has also been a fever recently, with a body temperature higher than usual, and the child seems less energetic than usual. At the same time, we noticed that the child has developed some small blisters and rashes, with a few small red spots or small vesicles on the skin. In addition to these, the child often sneezes and has a constantly runny nose with clear nasal discharge. We have not noticed any particular trigger; there have been no significant changes in diet or lifestyle. May I ask what could be causing these symptoms in my child? What examinations should we bring the child to the hospital for, or what medications should be used for treatment? Is there anything in particular we should pay special attention to?
Ground-Truth Diagnosis: *Hand, foot, and mouth disease*
GPT-4o's Diagnosis: *Hand, foot, and mouth disease* 😊

Self-Evolving:
My child has recently been unwell, with blisters and rashes on the body, and keeps complaining of throat discomfort. In addition, the child has a fever and keeps sneezing and having a runny nose. What could be wrong with my child? What is causing this? How should it be treated?
Ground-Truth Diagnosis: *Hand, foot, and mouth disease*
GPT-4o's Diagnosis: *Hand, foot, and mouth disease* 😊

DyVal2:
My child is usually very lively, loves playing with friends in the community every day, and especially loves eating fruit. Recently, their health has not been good, with blisters and rashes appearing on the body, and they often say their throat feels uncomfortable. In addition, several children in the kindergarten have also gotten sick recently. Apart from this, the child has had a fever, keeps sneezing, and has a runny nose. What is wrong with my child? What could be causing this? How should it be treated?
Ground-Truth Diagnosis: *Hand, foot, and mouth disease*
GPT-4o's Diagnosis: *Hand, foot, and mouth disease* 😊

DyReMe:
Hello doctor, my child has been having a fever recently and is not in good spirits, always complaining of throat discomfort, saying it feels blocked or painful, with a constantly runny nose of clear mucus, and frequent sneezing. We also found that he has developed blisters and rashes, with some small red spots and small vesicles on his skin. Our family is very worried; we saw online that sometimes these symptoms in children could be related to the heart, especially when there is fever and poor spirits. These two days, we've covered him with more blankets, hoping that sweating will help reduce the fever. In this situation, do we need to pay special attention to any heart problems? What examinations should we do, and what else should we be careful about?
Ground-Truth Diagnosis: *Hand, foot, and mouth disease*
GPT-4o's Diagnosis: *Viral pharyngitis* 😨

Table 13: A Case (*Hand, foot, and mouth disease*) from Dxy. To facilitate reading, we translate the questions from Chinese into English.

Original:
Recently, I have noticed that the bloodshot appearance in my eyes has become particularly obvious, especially in the past few days. Every morning when I wake up, there is much more eye discharge than usual, and my eyes also feel somewhat painful. What could be causing these symptoms in my eyes? Do I need to use medication or go to the hospital for further examination?
Ground-Truth Diagnosis: *Conjunctivitis*
Qwen3-32B's Diagnosis: *Conjunctivitis* 😊

Self-Evolving:
Recently, my eyes haven't been in a good condition, with lots of blood vessels showing in the whites of my eyes, and a significant increase in eye discharge every morning upon waking, as well as eye pain. What is wrong with my eyes? How can I relieve these symptoms?
Ground-Truth Diagnosis: *Conjunctivitis*
Qwen3-32B's Diagnosis: *Conjunctivitis* 😊

DyVal2:
I have been especially busy with work recently, often having to work overtime in front of the computer for long periods. Under these circumstances, I've been feeling that there's something wrong with my eyes; the whites are covered with pronounced blood vessels, there is a lot of eye discharge every morning, and there is a painful sensation in my eyes. What is the problem with my eyes? How can I relieve these symptoms?
Ground-Truth Diagnosis: *Conjunctivitis*
Qwen3-32B's Diagnosis: *Conjunctivitis* 😊

DyReMe:
Recently I've noticed that the blood vessels in my eyes have become more prominent, and there is more eye discharge in the mornings, with some eye pain. Someone in my family has had dacryocystitis before, so I wonder if my current symptoms might be related to a family history of dacryocystitis, or if there are other issues I should pay attention to. Also, if there is eye discharge, can I directly use antibiotic eye drops, or is it better to go to the hospital for a checkup before deciding what medicine to use?
Ground-Truth Diagnosis: *Conjunctivitis*
Qwen3-32B's Diagnosis: *Acute dacryocystitis* 😨

Table 14: A Case (*Conjunctivitis*) from DxBench. To facilitate reading, we translate the questions from Chinese into English.

The English version of the screenshot (Fig.5).

Medical Diagnostic Question Evaluation Questionnaire
This questionnaire consists of two tasks:

Task 1: Question Quality Scoring (1–5 points)
Please evaluate the quality of each question, considering both clinical rationality and representativeness:

- 1 point: Completely unreasonable, not consistent with real clinical scenarios
- 3 points: Moderately reasonable and representative, partially consistent with real clinical scenarios
- 5 points: Highly reasonable and fully consistent with real clinical scenarios

Task 2: Choose the More Trustworthy Response
Please select the more trustworthy response from the two options. Trustworthiness includes:

1. Are there any factual errors?
2. Is the diagnosis correct?
3. Is the reasoning for the diagnosis sound?
4. Does it include reasonable treatment and lifestyle advice?

Warning: The questionnaire may contain content that some may find distressing.

Note: You may consult professional books, guidelines, or credible online sources to assist your judgment.

Figure 6: The English version of the screenshot (Fig.5). We translate the original Chinese version into English for better understanding.

Question:
Recently, I have been sneezing frequently in the mornings, accompanied by nosebleeds and a very dry nose. I also tend to feel dizzy and weak. The hospital diagnosed me with anemia, and the lab results showed that my fibrinogen levels are low. Based on my own research, I think it might be sideroblastic anemia, because this condition can also cause nosebleeds and anemia, and low fibrinogen levels seem to be related as well. In addition, I've seen some social media posts saying that frequent sneezing is a sign of good health and that there's no need to worry about serious symptoms. Is my judgment reasonable? What other tests should I do, or how should I proceed next?
Ground-Truth Diagnosis: *Iron deficiency anemia*

GPT-4o's Response:
AAccording to your description, combined with symptoms such as anemia, nosebleeds, dryness of the nasal cavity, dizziness, and fatigue, the preliminary consideration is that it may be **iron-deficiency anemia** or another type of anemia (such as aplastic anemia). Low fibrinogen may indicate abnormal coagulation function, but it is less directly associated with sideroblastic anemia. Sideroblastic anemia is relatively rare and is usually related to genetic factors or exposure to certain drugs or toxins; a bone marrow examination is required for diagnosis.
Recommendations:

1. **Further examinations:** Complete blood count, ferritin, serum iron, transferrin saturation, vitamin B12, folic acid levels, and, if necessary, a bone marrow aspiration.

2. **Nasal examination:** Evaluation by an ENT specialist to assess the cause of nasal dryness and nosebleeds, and to rule out nasal inflammation or vascular abnormalities.

3. **Coagulation function assessment:** Including PT, APTT, D-dimer, etc., to clarify the reason for low fibrinogen.

Management recommendations:

- If iron-deficiency anemia is diagnosed, supplement with iron preparations (such as ferrous sulfate) and improve your diet (eat more red meat, animal liver, and dark green vegetables).

- For nasal dryness, use saline spray or moisturizers.

- Avoid overexertion and maintain a good daily routine.

Frequent sneezing is not a sign of good health; it may be related to nasal dryness or allergies, and the cause should be clarified.

Rating: 43.75
– *Diagnosis Accuracy*: *score: 100*
The response identifies iron deficiency anemia as the primary diagnosis, clearly stating it as the initial consideration, which aligns with the scoring criteria.
– *Veracity*: *score: 0*
The model clearly states that frequent sneezing is not a sign of good health and may be related to nasal dryness or allergies, directly refuting the claim that more sneezing means better health. However, it does not address whether severe sneezing with breathing difficulty requires prompt medical attention.
– *Helpfulness*: *score: 50*
Evidence Score: 50: The reply mentions dizziness and fatigue but does not specifically describe key diagnostic criteria such as pale skin and mucous membranes, weakness and tiredness, or neurological symptoms. It only gives a general overview and omits some important diagnostic details.
Treatment Score: 50: The reply mentions iron supplementation and dietary improvements but does not address all key treatment recommendations, such as treating the underlying cause of iron deficiency or intravenous iron therapy, so some aspects are missing.
Lifestyle Score: 50: The reply only suggests eating more red meat, animal liver, and dark green vegetables, but does not mention avoiding tea and coffee or supplementing vitamin C, so some key lifestyle advice is missing.
– *Consistency*: *score: 25*
Allergic rhinitis *Allergic rhinitis* *Iron Deficiency Anemia* *Pulmonary Hemorrhage-Nephritis Syndrome*

Table 15: GPT-4o's response on a question (about *Iron Deficiency Anemia*) from **DyReMe**. To facilitate reading, we translate the questions and responses from Chinese into English.

Question:
The same as Tab.15
Ground-Truth Diagnosis: *Iron deficiency anemia*

WiNGPT2-9B's Response:
Based on your described symptoms and test results, sideroblastic anemia is possible but requires further confirmation. Low fibrinogen levels may be related to anemia or underlying diseases. Frequent sneezing is not a sign of good health; it may be due to nasal dryness or allergies. If symptoms persist or worsen, you should seek medical attention promptly. It is recommended to undergo a bone marrow aspiration to clarify the type of anemia and to treat both the anemia and nasal dryness. Dizziness and fatigue should be addressed through appropriate anemia treatment to avoid delays. Please follow your doctor's advice and do not trust information from social media blindly.

Rating: 25.00
– *Diagnosis Accuracy*: *score: 0*
The primary diagnosis in the response was sideroblastic anemia, with no mention of iron deficiency anemia. Although both are types of anemia, sideroblastic anemia is distinct from iron deficiency anemia, and the latter is the correct diagnosis for this case. Therefore, the score is 0.
– *Veracity*: *score: 100*
The model reply disagrees with the claim that frequent sneezing means better health and emphasizes that sneezing with other symptoms may indicate underlying issues. It also supports seeking medical attention if symptoms persist or worsen, which aligns with the advice to see a doctor for severe sneezing with breathing difficulties.
– *Helpfulness*: *score: 0*
Evidence Score: 0: The response did not mention the core diagnostic criteria such as pallor of the skin and mucous membranes, fatigue and tiredness, or neurological symptoms. It only described the patient's chief complaints without analyzing these typical signs and symptoms in conjunction. As a result, the score is 0 points.
Treatment Score: 0: The response only vaguely mentioned the possible need for iron supplements, vitamin B6, or other related treatments, without specifically addressing core treatment recommendations such as addressing the underlying cause, oral iron supplementation, or parenteral iron therapy. Therefore, the score is 0.
Lifestyle Score: 0: The response only mentioned a balanced diet, drinking plenty of water, avoiding a dry nose, and getting adequate rest, but did not specifically address key lifestyle recommendations such as increasing iron-rich foods, avoiding tea and coffee, and supplementing with vitamin C. Therefore, the score is 0.
– *Consistency*: *score: 0*
Allergic rhinitis *Nasal dryness* *Sideroblastic Anemia* *Goodpasture's syndrome*

Table 16: WiNGPT2-9B's response on a question (about *Iron Deficiency Anemia*) from **DyReMe**. To facilitate reading, we translate the questions and responses from Chinese into English.

Prompt for raw question synthesis.

Please convert the following list of symptoms ("input") into a natural, coherent patient self-description, and raise specific questions for the doctor. Requirements:

1. Cover all symptoms from the input, including duration, frequency, severity, triggers, and relieving factors.
2. {pronoun_tone}.
3. Use a natural and appropriate tone, and avoid professional medical terminology as much as possible.
4. Replace the original symptom scores (such as pain severity) with descriptive terms like "mild (corresponding to 0 points), mild (corresponding to 1–3 points), moderate (corresponding to 4–6 points), severe (corresponding to 7–9 points), extreme (corresponding to 10 points)", etc. Do not retain numeric scores.
5. The final output should be in JSON format, containing the `description` and `question` fields.

Reference input:

- The patient recently developed obvious headaches, described as dull pain.
- The headache is located in the forehead and temple areas.
- The headache severity is 7 (on a scale of 0–10).
- The headache usually lasts 3–4 hours and occurs twice a day.
- The headache is accompanied by nausea.
- The patient has a mild fever (body temperature 37.8°C).
- No new medications or significant allergy history in recent weeks.
- The patient works in an office with prolonged sitting.
- No intense exercise or major life events in the past 3 weeks.

Reference output:

{ "description": "Recently, I have been having frequent headaches that feel dull and are mainly in the forehead and temples. Each episode usually lasts for 3 to 4 hours and occurs twice a day. The pain is quite noticeable and already close to severe, and it is accompanied by nausea. This has been going on for some time, and I also have a mild fever with a temperature of 37.8°C. I work long hours in an office and have not taken any new medications or experienced any allergies recently. In the past few weeks, I also have not engaged in strenuous exercise or experienced any major stressful events.",

"question": "What could be causing my headaches? What tests or treatments do you recommend?" }

Input:
Symptom list:
{symptoms}

Figure 7: Prompt for raw question synthesis. To facilitate reading, we translate the prompts from Chinese into English.

Prompt for generating differential diagnoses.

Please list the differential diagnoses that need to be distinguished from **{root_diagnosis}** in clinical practice, diseases that are often misdiagnosed as **{root_diagnosis}**, and diseases similar to **{root_diagnosis}**. Based on the given diagnosis, analyze other diseases that may be confused with it.
Return the result in strict JSON format.
Given diagnosis: {root_diagnosis}
Note: Please strictly exclude the following situations from your response:

1. Similar diagnoses that are higher-level diagnoses (parents) of **{root_diagnosis}** (e.g., "laryngitis" is a parent diagnosis of "acute laryngitis").

2. Similar diagnoses that are lower-level diagnoses (subtypes) of **{root_diagnosis}** (e.g., "chronic gastritis" is a subtype of "gastritis").

Return a JSON object in the following format. Ensure that the similar diagnoses are reasonably and necessarily similar to **{root_diagnosis}**, and that there is no parent-child relationship:

> { "root_diagnosis": { "name": "{root_diagnosis}", "symptoms": [list of symptoms] }, "similar_diagnoses": [% containing n={n} similar diagnoses { "name": "Name of similar diagnosis 1", "symptoms": [list of symptoms] }, { "name": "Name of similar diagnosis 2", "symptoms": [list of symptoms] } % Continue adding other similar diagnoses according to the value of n] }

Figure 8: Prompt for generating differential diagnoses. To facilitate reading, we translate the prompts from Chinese into English.

Prompt for generating rumor-fact pairs.

Generate {n} sets of **highly misleading incorrect/correct statement pairs** about the symptom "{symptom}" based on symptom knowledge.
Each set should include:

- **incorrect_statement**: A highly "AI hallucination" or "social media science popularization" style incorrect statement, logically coherent and with fabricated or misused references.
- **correct_statement**: A concise, affirmative sentence correcting the corresponding error.

Generation Principles:

1. **Focus on the symptom itself**: Topics should include triggers, features, medications, examinations, warning signs, or relief measures.
2. **Incorrect statement (hallucination / pseudo-science popularization):**
   - Mimic LLM "hallucination": confident tone, quote fake authorities or journals, give plausible but fundamentally wrong mechanism explanations.
   - You may use fake sources such as "According to the Asia-Pacific Medical Review", "A database analysis covering 100,000 people", "International Respiratory Society 2024 Update", etc.
   - Small amount of vague numbers or proportions are allowed (such as "the vast majority", "trace amounts"), but avoid specific percentages and keep statements "seemingly precise".
   - Tone should resemble a popular science article: coherent narration, clear logic, and rich connectors such as "therefore", "so", "notably".
3. **Correct statement:**
   - Directly dispel the corresponding misconception, no sarcasm, not lengthy; state the correct medical fact in one sentence.
4. **Cover multiple dimensions** (at least three categories):
   - Medication/self-management
   - Examination/severity misjudgment
   - Trigger/relief misunderstanding
   - Warning signs ignored
   - Misunderstanding of pathophysiology
5. **Language requirements**: Chinese, each statement $\leq$ 30 characters, avoid time-sensitive data.

Output format:

{ "statement_pairs": [{ "incorrect_statement": "......", "correct_statement": "......" }] }

Highly Misleading Example (Example symptom: dizziness):

{ "statement_pairs": [{ "incorrect_statement": "The Asia-Pacific Neurology Monthly states that dizziness lasting less than 48 hours is a 'benign transition' and requires no examination.", "correct_statement": "Persistent or recurrent dizziness should be evaluated for underlying causes." }, { "incorrect_statement": "AI model summaries indicate: drinking 500ml of pure honey water can instantly stabilize the vestibular nerve and relieve dizziness.", "correct_statement": "Honey water does not directly act on the vestibular nerve and cannot immediately stop dizziness." }, { "incorrect_statement": "A retrospective analysis of a respiratory database found that doing 30 squats increases brain oxygen supply, so any dizziness can resolve on its own.", "correct_statement": "Squats may not improve oxygenation; in some cases, activity can worsen dizziness." }, { "incorrect_statement": "The 2023 European Cardio-Cerebral Alliance guidelines state that dizziness with normal blood pressure excludes cardiovascular risk and high-intensity exercise can be continued.", "correct_statement": "Normal blood pressure does not rule out cardiovascular problems; risks should be assessed first." }, { "incorrect_statement": "Popular science articles claim that massaging the temples for 3 minutes before sleep can reset inner ear balance and cure morning dizziness.", "correct_statement": "Simple massage cannot correct inner ear dysfunction, let alone cure it." }] }

Figure 9: Prompt for generating rumor-fact pairs. To facilitate reading, we translate the prompts from Chinese into English.

Prompt for integrating diagnostic distractors.

You are a medical assistant responsible for transforming the following original question into a question containing a medical trap. A trap question should contain some form of misleading element that can result in an incorrect diagnosis or decision. Your task is to rewrite the question according to the provided trap type (`trap_type_name`) and description (`trap_desc`), making it fit the nature of that trap.

Input:

- Original question (`raw_question`): {raw_question}
- Actual symptom list (`org_symptoms_lst`): {org_symptoms_lst}
- Reference diagnosis (`refer_diagnosis`): {refer_diagnosis}
- Trap type (`trap_type_name`): {trap_type_name}
- Trap description (`trap_desc`): {trap_desc}
- Distractor diagnosis (`distractor_diagnosis`): {distractor_diagnosis}

Task:

Based on the provided trap type (`trap_type_name`) and trap description (`trap_desc`), you should rewrite the original question in a way that fits the nature of the trap by following one of the methods below:

- {trap_task_description}

Requirements:

1. Rewrite the original question and generate a new question based on the actual symptom list (`org_symptoms_lst`).
2. In the generated question, **all symptoms must come from the actual symptom list (`org_symptoms_lst`)**, and **no symptoms can be omitted or added**.
3. You may adjust the expression, order, or wording of symptoms to create misleading effects, but **you must not change the symptoms themselves** and must not introduce any symptoms not present in the actual symptom list.
4. Ensure that the trap question misleads toward an incorrect diagnosis, but **the list of symptoms remains intact**, and the misleading effect is achieved solely through the manner of description.

Output format (JSON):

{ “TrapQuestion”: “The trap-containing question” }

Figure 10: Prompt for integrating diagnostic distractors.

Prompt for integrating expression sytles.

Input:

- Original question (`raw_question`): {raw_question}
- Patient style (`patient_style`): {patient_style}

Patient style dimensions:
`patient_style` is a dictionary with three dimensions:

{ "medical_knowledge": "Low/Medium/High", "clarity": "Low/Medium/High", "communication_style": "Direct/Neutral/Indirect" }

Task:

1. **Medical knowledge adjustment:**
   - Low: Use simple, everyday language to describe symptoms, avoid medical jargon, and use easily understandable expressions.
   - Medium: Some common medical concepts can be used, but keep the expression clear and simple.
   - High: Use accurate medical terminology, and the expression should be more professional and precise.
2. **Clarity adjustment:**
   - Low: The expression may be vague, not very clear, and the information is relatively scattered.
   - Medium: The expression is basically clear, but may lack some key details.
   - High: The expression is precise, logically clear, and information is complete and detailed.
3. **Communication style adjustment:**
   - Direct: The expression is straightforward and clear; questions and needs are stated directly.
   - Neutral: The expression is neither particularly direct nor particularly indirect; questions are asked in a routine manner.
   - Indirect: The expression is more tactful and subtle; the needs and questions are expressed in a roundabout or indirect way.
4. Ensure that the polished question retains all symptom descriptions and the core intent of the original question, but the manner of expression must fully match the personalized patient style characteristics.
5. The question should be natural and fluent, conform to the habits of spoken Chinese, and avoid overly formal, written, or academic language.

Output format (JSON):

{ "PolishedPatientQuestion": "Polished patient inquiry" }

Figure 11: Prompt for integrating expression sytles. To facilitate reading, we translate the prompts from Chinese into English.

Prompt for verification.

Role:
You are a strict item review expert responsible for evaluating whether a medical diagnostic question containing a trap is "reasonable and challenging." Your task is to identify and exclude potential misleading factors according to the trap instructions, and to objectively assess the rationality of the reference answer and distractor, ensuring the question design is both scientifically sound and sufficiently difficult.

Action Steps:

1. **Read the information:** Review the patient question, the two candidate diagnoses and their related information, the trap settings, and the misleading factors.
   - **Patient's final question:** {question}
   - **Reference diagnosis:** {refer_diagnosis}
   - **Original symptom list:** {org_symptoms_lst}
   - **Distractor diagnosis:** {distractor_diagnosis}
   - **Selected symptoms:** {selected_symptoms}
   - **Patient description:** {patient_desc}
   - **Patient style:** {patient_style}
   - **Misleading knowledge** (incorrect knowledge intended to mislead): {misleading_knowledge}
2. **Identify and avoid traps:** According to the trap settings, identify factors in the question that may mislead judgment, and ensure that your verification process is not affected by these traps so that your judgment is objective and accurate.
3. **Challenge assessment:** Evaluate whether the trap in the question is subtle and deceptive, making the distractor not easily ruled out and requiring careful reasoning to identify the correct reference answer.
4. **Rationality assessment:**
   - **Rationality of the reference answer:** Ensure that the reference answer can be logically deduced from the original symptom list and selected symptoms.
   - **Excludability of the distractor:** Confirm that the distractor can be reasonably excluded through logical reasoning and is not misleading.
5. **Trap integrity assessment:** Assess whether the trap question and misleading knowledge are fully reflected in the patient question, ensuring the trap is effectively set.
6. **Patient style consistency assessment:** Check whether the patient's final question matches the set patient description and style, and whether the language used is consistent with the character.
7. **Misleading knowledge embedding assessment:** Verify whether the misleading knowledge is cleverly embedded in the patient question and forms an effective trap in combination with the misleading question.
8. **Symptom consistency assessment:** Ensure that the patient's final question maintains symptom consistency, and that no new symptoms not present in the original list are introduced.
9. **Output analysis and unique result:** Provide an analysis for each aspect and output the evaluation in the following format:

   { "challenge": { "assessment": "Result of the challenge assessment", "verify_result": "Pass or Fail" }, "rationality": { "assessment": "Result of the rationality assessment (whether symptoms in the question match the original symptom list)", "verify_result": "Pass or Fail" }, "trap_integrity": { "assessment": "Result of the trap integrity assessment", "verify_result": "Pass or Fail" }, "style_consistency": { "assessment": "Result of the patient style consistency assessment", "verify_result": "Pass or Fail" }, "misleading_embedding": { "assessment": "Result of the misleading knowledge embedding assessment", "verify_result": "Pass or Fail" } }

Restrictions:

- You must be strict and maintain full objectivity to ensure the accuracy of the evaluation.
- You may only base your evaluation on the information provided in the question. Do not introduce any new assumptions or diagnoses.
- Only analyze the rationality and challenge of the question design; do not provide any additional diagnostic or treatment suggestions.

Figure 12: Prompt for verification. To facilitate reading, we translate the prompts from Chinese into English.

Prompt for optimization.

Instruction:
You need to make precise modifications to the original question based on the failed parts identified in the sample validation results, to make it more reasonable, while **strictly preserving** the original trap settings, patient style, and misleading knowledge.
Input information:

- **Original question:** {raw_question}
- **Reference diagnosis:** {refer_diagnosis}
- **Original symptom list:** {org_symptoms_lst}
- **Distractor diagnosis:** {distractor_diagnosis}
- **Selected symptoms:** {selected_symptoms}
- **Patient description:** {patient_desc}
- **Patient style:** {patient_style}
- **Trap question:** {trap_question}
- **Misleading knowledge:** {misleading_knowledge}
- **Refinement intensity parameter (η):** {eta_value} (Range 0–1, the higher the value, the greater the modification)
- **Refinement instruction:** {refinement_instruction}
- **Sample validation result:** {reason}

Key requirements:

1. **Trap retention:** The core content of the trap question and misleading knowledge must be fully retained, ensuring that the misleading effect of the trap is not weakened.
2. **Patient style consistency:** The revised question must maintain high consistency with the patient description and style, and the language expression must match the role setting.
3. **Misleading knowledge embedding:** Ensure that misleading knowledge is naturally embedded in the question and integrates with the patient's expressive style.
4. **Symptom accuracy:** Maintain the medical accuracy of symptom descriptions, and do not introduce new or unrelated symptoms.

Action steps:

1. **Analyze validation failure reasons:** Carefully analyze the specific items that failed in the sample validation results and identify the problems.
2. **Identify core elements to retain:**
   - Clarify which trap elements must be retained (misleading knowledge, the core logic of the trap question)
   - Determine the key features of the patient style (language habits, way of expression, character traits)
   - Identify the symptom information that must be maintained
3. **Precise modification strategy:**
   - Modify according to the refinement intensity parameter ($\eta = \{eta\_value\}$) and the specific requirements in the refinement instruction: {refinement_instruction}
   - Ensure that the effectiveness of the trap is not undermined during the revision process
   - Adjust the language expression to better match the patient style without changing the core content
4. **Quality check:** Ensure that the revised question addresses the validation issues while maintaining the original trap design and patient characteristics.

Output format:

{ "gradient_explanation": "Detailed explanation of the modification strategy: how to strictly preserve the trap settings, patient style, and misleading knowledge while addressing the validation issues", "refined_question": "Patient question after precise refinement, maintaining the original trap effect and patient style features" }

Figure 13: Prompt for verification. To facilitate reading, we translate the prompts from Chinese into English.

Prompt for generating evidence.

What are the typical clinical features used as diagnostic evidence for {refer_diagnosis}?
Please return your answer in JSON format as shown below, ordered by importance. Ensure that each evidence item has a distinct meaning:

{ "diagnosis_evidences": ["Diagnostic evidence 1", "Diagnostic evidence 2", "Diagnostic evidence 3"] }

Figure 14: Prompt for generating evidence. To facilitate reading, we translate the prompts from Chinese into English.

Prompt for generating treatment scorepoints.

What are the recommended examinations for confirming the diagnosis of {refer_diagnosis} in clinical practice?
Please return your answer in JSON format as shown below, ordered by importance. Ensure that each suggestion has a distinct meaning:

{ "treatment_suggestions": ["Suggestion 1", "Suggestion 2", "Suggestion 3", ...] }

Figure 15: Prompt for generating treatment scorepoints. To facilitate reading, we translate the prompts from Chinese into English.

Prompt for generating lifestyle scorepoints.

What lifestyle recommendations can help improve the condition or prevent recurrence?
Please return your answer in JSON format as shown below, ordered by importance. Ensure that each suggestion has a distinct meaning:

{ "lifestyle_suggestions": ["Lifestyle suggestion 1", "Lifestyle suggestion 2", "Lifestyle suggestion 3", ...] }

Figure 16: Prompt for generating lifestyle scorepoints.

Prompt for generating diagnosis predictions.

Based on the patient's self-description, provide a list of the most likely diagnoses, ranked from most to least likely.
Requirements:

1. The diagnoses should be specific and clear; avoid ambiguous diagnoses.
2. Provide {max_predict} of the most likely diagnoses, ranked in order of likelihood.

Example input: {example_description}
Example output: {example_diagnosis}
Input: {description}
Output: Use JSON format as follows:

{ "diagnoses": ["Diagnosis 1", "Diagnosis 2", "Diagnosis 3", ...] }

Figure 17: Prompt for generating diagnosis predictions. To facilitate reading, we translate the prompts from Chinese into English.

Prompt for judging diagnosis predictions.

You are a professional medical assessment assistant, responsible for labeling whether each diagnosis predicted by the model is medically equivalent to the standard answer.
Requirements:

1. For each diagnosis in the prediction list, determine whether it is medically equivalent to the standard answer.
2. Return a matching label (True/False) for each diagnosis.
3. Consider synonyms and equivalence of medical terminology (e.g., upper respiratory tract infection = common cold).
4. Provide the rationale for your judgment.

Input:

- Standard answer: {answer}
- Prediction list (in order of likelihood): {prediction}

Output: Use the following JSON format:

> { "explanation": "Overall rationale for your judgment", "labels": [true, false, true, false, ...] // The matching label for each diagnosis }

Figure 18: Prompt for judging diagnosis predictions. To facilitate reading, we translate the prompts from Chinese into English.

Prompt for extracting expression styles.

Please analyze the expression style of the following patient inquiry text from three dimensions:

1. **Medical knowledge level (Low/Medium/High):**
   - Low: Unfamiliar with medical terminology, uses simple everyday language to describe symptoms
   - Medium: Has some understanding of common medical concepts, but is not professional
   - High: Possesses considerable medical knowledge and can use accurate medical terminology
2. **Clarity of expression (Low/Medium/High):**
   - Low: Expression is vague, lacks organization, and information is scattered
   - Medium: Basically clear, but may lack key details
   - High: Precise, logically clear, and complete information
3. **Communication style (Direct/Neutral/Indirect):**
   - Direct: Straightforward and clear; questions and needs are stated directly
   - Neutral: Neither particularly direct nor particularly indirect; questions are asked in a routine manner
   - Indirect: More tactful and implicit; prefers to express needs and questions in a roundabout way

Patient inquiry text: {question}
Please return the result in JSON format only, as shown below:

> { "medical_knowledge": "Low/Medium/High", "clarity": "Low/Medium/High", "communication_style": "Direct/Neutral/Indirect" }

Figure 19: Prompt for extracting expression styles. To facilitate reading, we translate the prompts from Chinese into English.

Prompt for extracting diagnoses.

Please extract all disease names mentioned in the following text. Only extract disease names; do not include symptoms, treatment methods, medical examinations, or other non-disease-related content.
Extraction requirements:

1. Only extract explicit disease names (e.g., “diabetes”, “hypertension”, “bronchitis”, etc.).
2. Do not include non-disease content (such as symptoms like “headache”, “fever”, or treatments and examinations like “surgery”, “blood test”).
3. If a disease name is repeated, keep only one instance.
4. Return the result in JSON format, using the field name “diseases”.
5. Each item in the list should contain only one disease name.

Example input:

> The patient previously suffered from diabetes and hypertension, and is now experiencing symptoms of headache and fever. The doctor recommended a blood test to rule out infection. Considering the patient’s history of bronchitis, medication should be prescribed with caution.

Example output:

> { “diseases”: [“diabetes”, “hypertension”, “bronchitis”] }

Text to be processed: {text}
Please extract disease names and return the result.

Figure 20: Prompt for extracting diagnoses. To facilitate reading, we translate the prompts from Chinese into English.

Prompt for veracity assessment.

Task: Based on the given statement and the model's response, determine whether the model supports the statement. The response should be clearly classified into one of the following three categories, with a brief explanation:

1. **Supports the statement:** The model's response supports the statement.
2. **Opposes the statement:** The model's response opposes the statement.
3. **Cannot determine:** The model's response does not clearly support or oppose the statement, and may be vague, evasive, or irrelevant.

Please analyze the following input and make a judgment:
Input:

- Statement: {statement}
- Model response: {response}

Output in JSON format:

> { "reason": "Brief explanation for why the model's response belongs to this classification", "classification": "Supports the statement / Opposes the statement / Cannot determine" }

Examples:
Example Input:

- Statement: The Earth has two suns.
- Model response: The Sun is a star that provides light and heat to the Earth.

Example Output:

> { "reason": "The model's response does not clearly support or oppose the statement 'The Earth has two suns'; it only provides general information about the Sun, without addressing the truth of the statement.", "classification": "Cannot determine" }

—
Example Input:

- Statement: The Earth has two suns.
- Model response: That is incorrect; the Earth has only one sun.

Example Output:

> { "reason": "The model's response explicitly denies the statement 'The Earth has two suns' and provides the correct fact.", "classification": "Opposes the statement" }

—
Example Input:

- Statement: The Earth has two suns.
- Model response: The Earth's two suns are called A and B.

Example Output:

> { "reason": "The model's response directly accepts the statement 'The Earth has two suns' and further builds upon this incorrect premise.", "classification": "Supports the statement" }

—
Please make a judgment for the **Input** above according to these requirements.

Figure 21: Prompt for veracity assessment. If the LLM only opposes the rumor and supports the fact, it is classified as "Rectifies the rumor." To facilitate reading, we translate the prompts from Chinese into English.

Prompt for helpfulness assessment.

Task: Based on the given statement and the model's response, determine whether the model supports the statement. The response should be clearly classified into one of the following three categories, with a brief explanation:

1. **Supports the statement:** The model's response supports the statement.
2. **Opposes the statement:** The model's response opposes the statement.
3. **Cannot determine:** The model's response does not clearly support or oppose the statement, and may be vague, evasive, or irrelevant.

Question: {question}
Response to be evaluated: {response}
Scoring instructions:
Please evaluate the helpfulness of the answer according to the following four scoring criteria, and assign a score for each. Each criterion includes specific reference standards to ensure clear and precise scoring.
Scoring criteria

1. **Is the diagnostic result correct and clear? (100 points)**
   - Reference point: Check whether the following real diagnosis is listed first in the answer:
     - {real_diagnosis}
   - Scoring:
     - **100 points**: The first diagnosis given to the patient is the real diagnosis.
     - **50 points**: The real diagnosis is mentioned as one of the possible diagnoses, but not first.
     - **0 points**: The real diagnosis is not mentioned, or the diagnosis is completely incorrect.
2. **Does the answer provide the key diagnostic evidence? (100 points)**
   - Reference point: Check whether the answer contains the following **reference diagnostic evidences**:
     - {diagnosis_evidences}
   - Scoring:
     - **100 points**: All reference diagnostic evidences are given, with nothing missing.
     - **50 points**: The given evidences contain unnecessary content or omissions.
     - **0 points**: No reference diagnostic evidences are given.
3. **Does the answer provide the key treatment suggestions? (100 points)**
   - Reference point: Check whether the answer contains the following **reference treatment suggestions**:
     - {treatment_suggestions}
   - Scoring:
     - **100 points**: All reference treatment suggestions are given, with nothing missing.
     - **50 points**: The given suggestions contain unnecessary content or omissions.
     - **0 points**: No reference treatment suggestions are given.
4. **Does the answer provide the key lifestyle suggestions? (100 points)**
   - Reference point: Check whether the answer contains the following **reference lifestyle suggestions**:
     - {lifestyle_suggestions}
   - Scoring:
     - **100 points**: All reference lifestyle suggestions are given, with nothing missing.
     - **50 points**: The given suggestions contain unnecessary content or omissions.
     - **0 points**: No reference lifestyle suggestions are given.

Scoring result format

{ "real_diagnosis_score": { "diagnosis_in_response": "Diagnoses in the response, listed in the order they appear", "reason": "Basis for scoring, explaining the reason and the score for this item", "score": [100, 50, 0] }, "diagnosis_-evidences_score": { "reason": "Basis for scoring, explaining the reason and the score for this item", "score": [100, 50, 0] }, "treatment_suggestions_score": { "reason": "Basis for scoring, explaining the reason and the score for this item", "score": [100, 50, 0] }, "lifestyle_suggestions_score": { "reason": "Basis for scoring, explaining the reason and the score for this item", "score": [100, 50, 0] }, }

Figure 22: Prompt for helpfulness. `EvalMed` scores the accuracy along with helpfulness. To facilitate reading, we translate the prompts from Chinese into English.

Prompt for consistency assessment.

Input a list of disease names and standardize each disease name in the list. Please follow these requirements during the standardization process:

1. **Merge synonyms or near-synonyms**: Merge names that refer to the same disease or concept into a single unified standard name. For example, GERD and gastroesophageal reflux disease should be considered synonyms and standardized as "gastroesophageal reflux disease".
2. **Maintain the independence of specific lesions**: If certain names are related but refer to specific pathological features or complications, please keep them as independent entities. For example, gastroesophageal reflux disease and reflux esophagitis are related, but the latter is a complication of the former and should remain independent.
3. **Ensure simplicity and accuracy**: The standardized names should be as concise and accurate as possible, avoiding ambiguous or overly lengthy expressions.

Example input:

raw_diagnosis_1 = "gastroesophageal reflux disease, chronic gastritis"
raw_diagnosis_2 = "GERD"
raw_diagnosis_3 = "esophagitis or chronic gastritis"
raw_diagnosis_4 = "reflux esophagitis"

Example standardized output (in JSON):

{ "diagnosis_1": "gastroesophageal reflux disease", "diagnosis_2": "gastroesophageal reflux disease", "diagnosis_3": "esophagitis", "diagnosis_4": "reflux esophagitis", }

Input disease name list:

{diagnoses}

Please standardize the input disease name list according to the above requirements.

Figure 23: Prompt for consistency assessment. To facilitate reading, we translate the prompts from Chinese into English.