A statistical framework for evaluating repeatability and reproducibility of large language models in diagnostic reasoning

Cathy Shyr[1*], Boyu Ren[2], Chih-Yuan Hsu[3], Rory J. Tinker[4], Rizwan Hamid[4], Adam Wright[1], Bradley A. Malin[1], Hua Xu[5]

[1*]Department of Biomedical Informatics, Vanderbilt University Medical Center, 2525 West End Avenue, Nashville, 37203, TN, USA.
[2]Laboratory for Psychiatric Biostatistics, McLean Hospital, 115 Mill Street, Belmont, 02478, MA, USA.
[3]Department of Biostatistics, Vanderbilt University Medical Center, 2525 West End Avenue, Nashville, 37203, TN, USA.
[4]Department of Pediatrics, Vanderbilt University Medical Center, 2200 Children's Way, Nashville, 37232, TN, USA.
[5]Department of Biomedical Informatics and Data Science, Yale School of Medicine, 100 College Street, New Haven, 06510, CT, USA.

*Corresponding author(s). E-mail(s): cathy.shyr@vumc.org;
Contributing authors: bor158@mail.harvard.edu; chih-yuan.hsu@vumc.org; rory.tinker@vumc.org; rizwan.hamid@vumc.org; adam.wright@vumc.org; b.malin@vumc.org; hua.xu@yale.edu;

Abstract

A major concern in applying large language models (LLMs) to medicine is their output variability, as they can generate different responses even when the input prompt, model architecture, and parameters remain the same. In this study, we present a statistical framework for evaluating LLM consistency by quantifying the repeatability and reproducibility of outputs in diagnostic reasoning. The framework captures both semantic variability, reflecting differences in the clinical meaning of outputs, and internal variability, reflecting differences in token-level generation behavior. These dimensions are critical in medicine, where subtle shifts in meaning or model reasoning may influence clinician interpretation and decision-making. We apply the framework across multiple LLMs using validated diagnostic prompts on standardized medical exam vignettes and real-world rare

disease cases from the Undiagnosed Diseases Network. We find that LLM consistency depends on the model, prompt, and complexity of the patient case, and is generally not correlated with diagnostic accuracy. This highlights the need for case-by-case assessment of output consistency to ensure reliability in clinical applications. Our framework can support model and prompt selection to promote more reliable use of LLMs in medicine.

Keywords: repeatability, reproducibility, large language model, artificial intelligence, diagnosis

1 Introduction

Large language models (LLMs) are artificial intelligence (AI) systems capable of learning complex language patterns and generating human-like text. They have demonstrated promising performance across a range of clinical applications, including drafting hospital discharge summaries [1], responding to patient messages [2], and providing diagnostic support [3]. Among these, diagnostic support has emerged as a particularly promising area, with recent studies highlighting strong LLM performance in diagnostic interpretation and clinical decision-making [4–7]. Although prior evaluations focused primarily on assessing diagnostic accuracy, other key dimensions critical for clinical adoption, such as consistency, remain underexplored.

Consistency, or a model's ability to produce similar outputs across repeated runs under the same or similar conditions, is critical for the clinical reliability of LLMs [8]. However, LLMs are probabilistic, such that they can generate different outputs even when the input prompt, model architecture, and parameters remain identical. This variability can raise concerns in diagnostic settings, as the same patient case may yield divergent diagnostic suggestions, potentially undermining clinician trust and limiting the model's utility for decision support [9, 10]. To this end, the U.S. Food and Drug Administration (FDA)'s draft guidance on AI-enabled medical software recommends evaluation of two properties: repeatability, the agreement of outputs under identical conditions, and reproducibility, the agreement under pre-specified, different conditions [11]. These principles align with broader calls by medical journals to assess variability as a core component of clinical AI evaluation [12].

In this study, we present a statistical framework for evaluating the repeatability and reproducibility of LLM outputs in diagnostic reasoning tasks. Our framework assesses both semantic and internal consistency. Semantic consistency captures the stability in the *meaning* of outputs across repeated runs or prompt variations. By contrast, internal consistency quantifies *variability in token-level generation behavior*. These aspects are particularly important in clinical settings, where small shifts in meaning or model reasoning may influence clinician interpretation and downstream decision-making. In comparison to prior studies that focused on assessing the *correctness* of LLM output (e.g., diagnostic accuracy [3, 6] or hallucination detection [13–15]), our approach systematically quantifies output *variability* to provide a complementary assessment of LLM reliability.

We apply this framework to a range of clinical scenarios, including standardized benchmark questions from the U.S. Medical Licensing Examination (USMLE) [16] and real-world rare disease cases from the NIH Undiagnosed Diseases Network (UDN) [17]. Using multiple LLMs and five validated diagnostic reasoning prompts [3], we systematically evaluate output consistency across models and clinical contexts. Our findings reveal variability in LLM behavior that is not captured by accuracy metrics alone, with implications for model selection, clinical implementation, and regulatory oversight.

More broadly, the framework offers a generalizable approach for assessing LLM consistency in clinical applications. By quantifying both semantic and internal variability, it creates an opportunity for more transparent and clinically meaningful evaluation of LLMs. As interest grows in defining standards for clinical AI development and oversight, this framework provides a structured, reproducible method for characterizing output variability.

2 Methods

2.1 Definitions of Repeatability and Reproducibility

Motivated by the FDA's draft guidance on AI-enabled medical software, which recommends evaluation of both repeatability and reproducibility in AI systems [11], our framework defines and operationalizes four metrics: 1) Semantic Repeatability, 2) Internal Repeatability, 3) Semantic Reproducibility, and 4) Internal Reproducibility (**Table** 1). Repeatability is defined by the FDA as *"the closeness of agreement of repeated measurements taken under the same conditions."* In our framework, this corresponds to generating output across repeated runs using the same LLM, prompt, and generation parameters (e.g., temperature and top-k), and measuring the agreement across these outputs. In contrast, reproducibility is defined as *"the closeness of agreement of repeated measurements taken under different, pre-specified conditions."* In our framework, this corresponds to generating output across repeated runs using the same LLM, parameter setting, and different, pre-specified diagnostic reasoning prompts.

2.2 Diagnostic Reasoning Prompts

We evaluate LLM repeatability and reproducibility across five diagnostic reasoning prompts developed and validated by Savage *et al.* [3], each designed to elicit a distinct reasoning strategy. We selected these prompts because they reflect distinct cognitive strategies used in clinical practice and were validated to produce interpretable and accurate model output. The prompts include traditional chain-of-thought (CoT) reasoning, differential diagnosis CoT, intuitive reasoning CoT, analytic reasoning CoT, and Bayesian reasoning CoT (**Table 1**). The full prompts used in this study are provided in **Supplementary Note 1**.

2.3 Data Sources

To evaluate the repeatability and reproducibility of LLM diagnostic reasoning across diverse clinical scenarios, we selected two distinct datasets: a standardized benchmark for general medical knowledge and a real-world rare disease cohort (**Table 1**).

MedQA Dataset. We used the publicly available MedQA dataset, a benchmark for medical question answering based on diagnostic clinical vignettes from the U.S. Medical Licensing Examination (USMLE). Following Savage et al. [3], we used the same set of 518 questions for evaluation. These questions were reformulated from multiple choice to free response format and focus on Step 2 and Step 3 cases that require clinical reasoning rather than rote memorization. An example is shown in **Table 2**. All questions are publicly available and provided in **Supplement Data 1** of Savage et al. [3].

Undiagnosed Diseases Network (UDN) Dataset: To address concerns that public datasets like MedQA may have been seen during LLM pre-training, we additionally used a non-public dataset of 90 diagnostically challenging rare disease cases from the Vanderbilt UDN. These cases consist of multi-paragraph summaries of each patient's medical history and diagnostic workup. An example is shown in **Table 2**. Written informed consent was obtained by all patients. This study was approved by the Vanderbilt University Medical Center Institutional Review Board (#172005).

2.4 Models

We performed our evaluation using three LLMs selected to represent a mix of commercial and open-source systems, model sizes, and intended use cases (**Table 1**). **ChatGPT-4 (OpenAI)** is a state-of-the-art commercial LLM previously shown to emulate clinical reasoning while maintaining high diagnostic accuracy by Savage *et al.* on the same set of 518 MedQA questions [3]. To this end, we selected this model to evaluate not only its repeatability and reproducibility, but also to explore whether these properties are correlated with the accuracy metrics reported in Savage *et al.* [3]. **ChatGPT-4o-mini (OpenAI)** is a smaller, more cost-efficient variant that allows us to assess repeatability and reproducibility under a practical, lightweight setting. **Llama 3.2-1B (Meta)** is an open-source, lightweight model that provides an efficient alternative to commercial systems. Its inclusion helps evaluate the feasibility of reproducibility using lightweight models in resource-limited settings.

2.5 Evaluation Setting

For each of the five prompts, we evaluated 518 MedQA cases and 90 UDN cases using each of the three LLMs, with $R = 100$ independent runs per prompt-case-model combination, totaling 912,000 generations. All models were accessed via the Microsoft Azure OpenAI API. To ensure patient privacy, all runs involving UDN cases were performed using secure Azure OpenAI instances. We set the temperature to 0.5 and used a top-k of 30 to balance output diversity with coherence by sampling from the most probable tokens. These settings were selected to strike a balance between determinism and diversity to enable meaningful assessment of both repeatability and reproducibility.

Table 1 Overview of evaluation framework. USMLE = U.S. Medical Licensing Examination.

Metric	Definition	Interpretation
Semantic Repeatability	Consistency in the meaning of outputs across repeated runs with the same LLM, parameters, and prompts	How consistently the model output conveys the same meaning across repeated runs under identical conditions
Internal Repeatability	Consistency in the token-level probability distributions across repeated runs with the same LLM, parameters, and prompts	How stable the model's internal sampling behavior is across repeated runs under identical conditions
Semantic Reproducibility	Consistency in the meaning of outputs across repeated runs with the same LLM, parameters, and different, pre-specified prompts	How robust the model's semantic output is to different prompts for the same clinical case
Internal Reproducibility	Consistency in the token-level probability distributions across repeated runs with the same LLM, parameters, and different, pre-specified prompts	How stable the model's internal sampling behavior is to different prompts for the same clinical case
Prompt Name	**Prompt**	**Reasoning Paradigm**
Traditional CoT	Provide a step-by-step deduction that identifies the correct response.	General logical reasoning aimed at identifying the correct answer.
Differential Diagnosis CoT	Use step-by-step deduction to create a differential diagnosis and then determine the correct response.	Clinical reasoning incorporating differential diagnosis before narrowing it down to the final diagnosis.
Intuitive Reasoning CoT	Use symptoms, signs, and lab associations to step-by-step deduce the correct response.	Heuristic reasoning grounded in clinical pattern recognition and associations.
Analytic Reasoning CoT	Use analytic reasoning to deduce the pathophysiology and step-by-step identify the diagnosis.	Mechanistic reasoning focused on underlying biological or physiological processes.
Bayesian Reasoning CoT	Use Bayesian inference to create a prior, update with new information, and determine the diagnosis.	Probabilistic reasoning involving dynamic updating of diagnostic likelihoods based on new evidence.
Data	**MedQA[16] USMLE ($n = 518$)**	**Rare Disease Cases ($n = 90$)**
Source	Public benchmark	NIH Undiagnosed Diseases Network
Content	Standardized exam-style vignettes	Real-world patient cases
Completeness	Fully specified questions	Often incomplete or open-ended
Intended Use	Medical reasoning evaluation	Rare disease diagnosis and evaluation
Clinical Context	General medical knowledge	Atypical, complex, and heterogeneous presentations
Realism	Synthetic and often idealized	High-fidelity, real clinical data
LLM	**Type**	**Rationale**
ChatGPT-4	Commercial	Shown to emulate clinical reasoning without loss in diagnostic accuracy [3]
ChatGPT-4o-mini	Commercial	Lightweight and cost efficient
Llama 3.2-1B	Open Source	Lightweight and efficient

Table 2 Example of USMLE and UDN clinical cases. USMLE = U.S. Medical Licensing Examination; UDN = Undiagnosed Diseases Network.

USMLE Case	UDN Case
A 55-year-old man comes to the physician because of a 6-week history of tingling pain in the sole of his right foot when he raises it above chest level during exercises. He reports that he started exercising regularly 2 months ago and that his right calf cramps when he uses the incline feature on the treadmill, forcing him to take frequent breaks. The pain completely disappears after resting for a few minutes. He has an 8-year history of type 2 diabetes mellitus. He has smoked two packs of cigarettes daily for 34 years. His only medication is metformin. His pulse is 82/min, and blood pressure is 170/92 mm Hg. Straight leg raise test elicits pallor and tingling pain in the right foot. There is no pain in the back. His muscle strength is normal. Femoral pulses are palpable; right pedal pulses are absent.	**"One-liner":** 6-year-old male with short stature, developmental delay, dysmorphic facial features, cryptorchidism, and pulmonary valve stenosis. **Category of Primary Condition:** Neurology **Narrative Summary:** At birth, noted to have generalized hypotonia, poor suck, and distinct facial features. Required NG feeding for the first two weeks of life. Developmental delays were evident from infancy. Rolled at 8 months, sat independently at 15 months, walked at 3 years. Now 6 years old, speaks only a few single words. Formal developmental assessment confirmed global developmental delay. Behavioral profile includes repetitive hand movements, mild incoordination, and attention difficulties. **Medical history notable for:** Pulmonary valve stenosis (stable) Bilateral cryptorchidism (surgically repaired) Recurrent otitis media (tubes x2) Severe constipation (daily laxatives) **Percentiles:** Height: <3rd percentile Weight: 5th percentile Head circumference: 10th percentile **Facial Features and Measurements:** Hypertelorism (intercanthal distance: 3.6 cm, 95th percentile) Down-slanting palpebral fissures Ptosis Low-set, posteriorly rotated ears (Right: 4.2 cm, Left: 4.3 cm; both ¡3rd percentile) Broad/webbed neck Shield chest with widely spaced nipples High anterior hairline **Family History:** Father: Short stature and had a heart murmur in childhood, never fully evaluated Paternal grandmother: Described as having a similar facial appearance Mother: Healthy **Prior Genetic Testing:** Chromosomal Microarray: Normal Fragile X testing: Negative Whole Exome Sequencing (trio): VUS in PTPN11, currently under review Karyotype: 46,XY **Known Prior Evaluations:** Brain MRI (2023): Normal Echocardiogram: Mild-to-moderate pulmonary valve stenosis, stable Audiology: Mild conductive hearing loss (likely secondary to otitis media)

2.6 Statistical Framework for Evaluating LLM Repeatability and Reproducibility

We formalize repeatability and reproducibility across two complementary dimensions, semantic and internal, resulting in a total of four metrics: (1) Semantic Repeatability, (2) Internal Repeatability, (3) Semantic Reproducibility, and (4) Internal Reproducibility (**Table 1**). Let X denote the input prompt and $\mathcal{V}$ the LLM's output vocabulary (i.e., set of all possible output tokens). For a given LLM and prompt, we generated R independent runs using fixed generation settings. Let $Y_{r,i} \in \mathcal{V}$ denote the output token generated at position $i \quad (i = 1, \ldots, L_r)$ in run $r \quad (r = 1, \ldots, R)$, where L_r is the output length for run r. The full output sequence for run r is denoted as $Y_{r,1:L_r}$.

2.6.1 Semantic Repeatability

Semantic repeatability measures the consistency in the *meaning* of the LLM's output across repeated runs under identical conditions. To quantify this, we define an embedding function $\mathcal{E} : \mathcal{V} \to \mathbb{R}^d$ that maps the LLM's output to a $d-$dimensional numerical vector. In this study, we use `MedEmbed-Large-v1` [18] as our embedding function, as it was specifically trained on clinical text and is well-suited for clinical applications. Importantly, our framework is agnostic to the choice of embedding model and can accommodate any embedding function appropriate for the task at hand. For each run $r = 1, \ldots, R$, we compute the semantic embedding vector $e_r := \mathcal{E}(Y_{r,1:L_r}) \in \mathbb{R}^d$. We define **Semantic Repeatability Score** $\bar{S}^{\text{Rpt}}$ as the average pairwise cosine similarity between all embeddings, with larger values of $\bar{S}^{\text{Repeatability}}$ indicating higher semantic repeatability. Stopwords (e.g., "the," "an") are removed prior to generating the embeddings to reduce noise from non-informative tokens.

Semantic Repeatability (Rpt) Score (Larger = More Reproducible)

$$\bar{S}^{\text{Rpt}} = \frac{2}{R(R-1)} \sum_{1 \le r < s \le R} \cos(e_r, e_s), \qquad -1 \le \bar{S}^{\text{Rpt}} \le 1.$$

2.6.2 Internal Repeatability

Internal repeatability measures the consistency in the LLM's output at the token level across repeated runs under identical conditions. In other words, it measures the variability of the model's internal token-level generation behavior. Lower internal repeatability indicates greater variability in the LLM's token-level generation, suggesting less stability in its behavior under identical conditions. To quantify internal repeatability, we let $f_i(X, Y_{r,1:i-1}) =: z_{r,i} \in \mathbb{R}^{|\mathcal{V}|}$ denote the vector of unnormalized logits at position i in run r given the input prompt X and previously generated output tokens $Y_{r,1:i-1}$. These logits are transformed into a probability distribution via

the temperature-scaled softmax function,

$$\pi_{r,i}(v) = \frac{\exp\{z_{r,i}(v)/T\}}{\sum_{v' \in \mathcal{V}} \exp\{z_{r,i}(v')/T\}}, \quad v \in \mathcal{V},$$

where T denotes the temperature. To focus on the subset of tokens most likely to be sampled and reduce the influence of low-probability tokens in the tails, we truncate this distribution to the top-k most probable tokens at each position:

$$\tilde{\pi}_{r,i}(v) = \begin{cases} \pi_{r,i}(v) \Big/ \sum_{u \in K_{r,i}} \pi_{r,i}(u), & v \in K_{r,i} \\ 0, & v \notin K_{r,i} \end{cases}$$

where $K_{r,i} \subset \mathcal{V}$ is the set of top-k tokens at position i in run r. To quantify token-level variability at position i, we compute the entropy of the LLM's output distribution $\tilde{\pi}_{r,i}(\cdot)$:

$$H_{r,i} = -\sum_{v \in K_{r,i}} \tilde{\pi}_{r,i}(v) \log \tilde{\pi}_{r,i}(v),$$

and average over positions in run r to obtain

$$H_r = \frac{1}{L_r} \sum_{i=1}^{L_r} H_{r,i}.$$

Finally, we define the **Internal Repeatability Score** $\bar{H}^{\text{Rpt}}$ as the average entropy across runs. Because entropy reflects variability, with higher values indicating greater uncertainty, we multiply by -1 so that larger values of the Internal Repeatability Score correspond to greater repeatability. As with semantic repeatability, stopwords are excluded to focus on meaningful variation.

Internal Repeatability (Rpt) Score (Larger = More Reproducible)

$$\bar{H}^{\text{Rpt}} = -\frac{1}{R} \sum_{r=1}^{R} H_r, \qquad -\log_2 k \leq \bar{H}^{\text{Rpt}} \leq 0.$$

2.6.3 Semantic Reproducibility

Semantic reproducibility measures the consistency in the *meaning* of the LLM's output across repeated runs using different, pre-specified prompts for the same clinical case, under otherwise identical conditions. In our study, we used $P = 5$ diagnostic reasoning prompts (**Table 1**). For each prompt $p = 1, \ldots, P$, we generate R independent runs and compute the semantic embeddings for each run using the embedding function $\mathcal{E}$ described in **Section 2.6.1**. Let $e_r^{(p)} = \mathcal{E}(Y_{r,1:L_r}^{(p)})$ denote the embeddings of run r under prompt p. We define the average embedding for each prompt as follows:

$$\bar{e}^{(p)} = \frac{1}{R}\sum_{r=1}^{R} e_r^{(p)}, \qquad p = 1, \ldots, P.$$

We define the **Semantic Reproducibility Score** $\bar{S}^{\mathrm{Rpd}}$ as the average pairwise cosine similarity between the mean embeddings produced by each prompt. Higher values of $\bar{S}^{\mathrm{Rpd}}$ indicate greater reproducibility of semantic meaning across different prompts.

Semantic Reproducibility (Rpd) Score (Larger = More Reproducible)

$$\bar{S}^{\mathrm{Rpd}} = \frac{2}{P(P-1)} \sum_{1 \le p < q \le P} \cos(\bar{e}^{(p)}, \bar{e}^{(q)}), \qquad -1 \le \bar{S}^{\mathrm{Rpd}} \le 1.$$

2.6.4 Internal Reproducibility

Internal reproducibility measures the stability of the LLM's *internal token-level sampling behavior* across repeated runs with different, pre-specified prompts for the same clinical case, under otherwise identical conditions. In other words, it evaluates how robust the model's token-level output distributions are to variation in prompts for the same clinical case. For each prompt $p = 1, \ldots, P$, we generate R independent runs and compute the average entropy. Let $H_r^{(p)}$ denote the mean entropy of the top-k output tokens in run r under prompt p. We define the average entropy for each prompt as:

$$\bar{H}^{(p)} := \frac{1}{R}\sum_{r=1}^{R} H_r^{(p)}, \qquad p = 1, \ldots, P.$$

We then compute the **Internal Reproducibility Score** $\bar{H}^{\mathrm{Rpd}}$ as the average absolute difference in mean entropy across all prompt pairs. We multiply the score by -1 so that larger values correspond to greater reproducibility.

Internal Reproducibility (Rpd) Score (Larger = More Reproducible)

$$\bar{H}^{\mathrm{Rpd}} = -\frac{2}{P(P-1)} \sum_{1 \le p < q \le P} \left| \bar{H}^{(p)} - \bar{H}^{(q)} \right|, \qquad -\log_2 k \le \bar{H}^{\mathrm{Rpd}} \le 0.$$

3 Results

3.1 Repeatability

Overall, both semantic and internal repeatability varied across prompts, LLMs, and datasets (**Figure 1**). For ChatGPT-4, the Bayesian CoT prompt resulted in higher semantic repeatability on both the USMLE and UDN cases, suggesting that probabilistic reasoning may elicit more stable outputs. In contrast, internal repeatability

scores were relatively consistent across models, prompts, and datasets. An exception is ChatGPT-4o-mini, where the Traditional CoT and Bayesian CoT prompts resulted in substantially lower internal repeatability. Compared to the USMLE cases, repeatability scores for UDN cases varied less across prompts, suggesting that the clinical ambiguity and complexity of real-world cases may constrain the LLM's output space and reduce prompt sensitivity due to more stable reasoning patterns.

3.2 Reproducibility

Reproducibility also varied by LLM and dataset (**Figure 2**). On the USMLE cases, ChatGPT-4o-mini generally achieved the highest internal reproducibility, indicating consistent token-level generation behavior across different prompts. In contrast, Llama 3.2-1B achieved the highest semantic reproducibility, suggesting greater robustness in the overall meaning of its outputs across different prompts. For the UDN cases, internal reproducibility was relatively uniform across all three models. However, Llama 3.2-1B outperformed the others in semantic reproducibility, indicating that its responses were more stable semantically in complex, real-world cases. These findings suggest that lightweight models like Llama 3.2-1B may exhibit stronger semantic robustness to prompt variation in complex clinical cases, possibly because their reduced capacity limits overfitting to prompt-specific patterns.

3.3 Relationship among repeatability, reproducibility, and diagnostic accuracy

To assess whether consistency is associated with diagnostic performance, we analyzed the relationship between repeatability and reproducibility scores for ChatGPT-4 on the USMLE cases stratified by diagnostic accuracy based on published ground-truth labels [3]. Briefly, these ground-truth labels were created by physicians who manually reviewed ChatGPT-4's output for diagnostic accuracy (c.f. Supplement 1 in [3]). As shown in **Figure 3**, across four of five prompts, traditional CoT, differential diagnosis CoT, analytic CoT, and Bayesian CoT, there was no statistically significant difference in repeatability scores between correctly and incorrectly diagnosed cases. For the intuitive CoT prompt, however, higher internal repeatability scores were positively associated with diagnostic accuracy (p-value < 0.001). This suggests that for certain reasoning paradigms, greater token-level consistency may correlate with improved accuracy. **Figure 4** shows reproducibility scores for the same cases. There was no statistically significant association between reproducibility and diagnostic accuracy, suggesting that accurate diagnoses do not necessarily correlate with consistent outputs across different prompts. Taken together, these findings highlight a distinction between generating an accurate diagnosis and generating that diagnosis *consistently*.

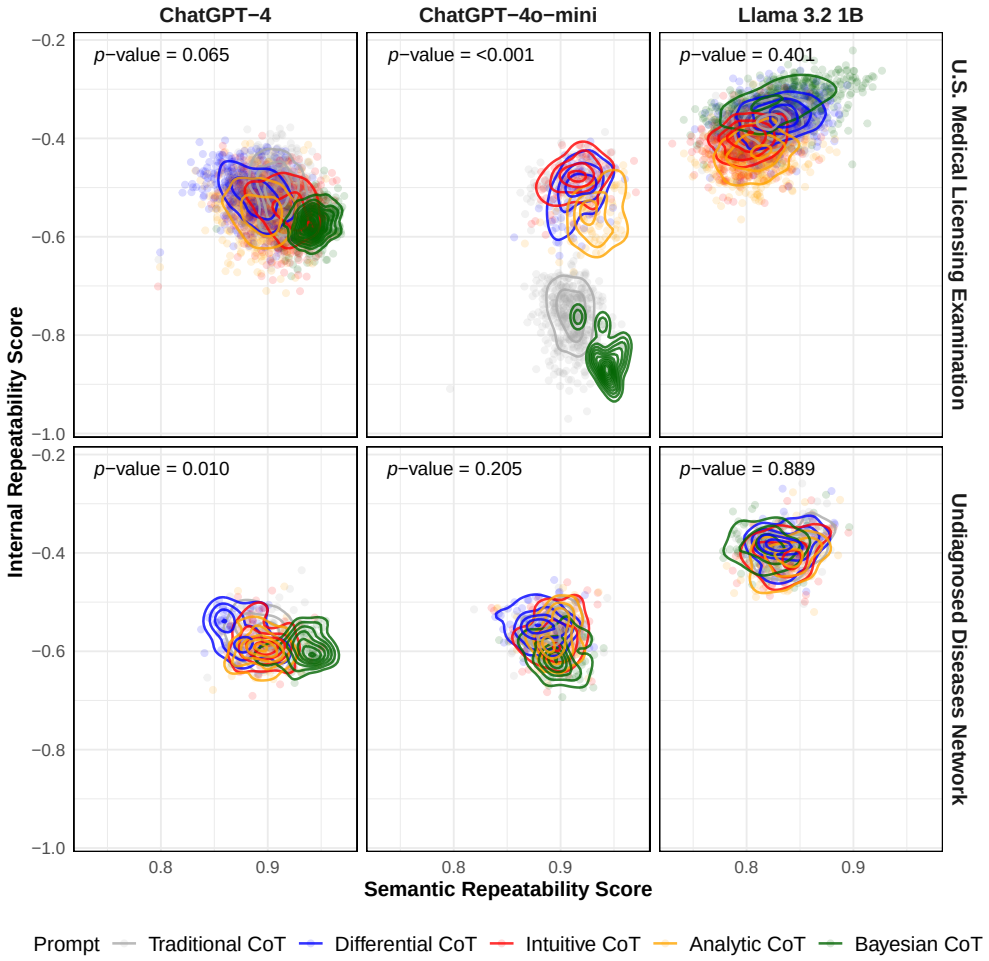

Fig. 1 Semantic and internal repeatability scores by model, dataset, and prompt. Each point represents a single clinical case evaluated under a specific prompt, plotted by its average semantic repeatability score (x-axis) and internal repeatability score (y-axis). Contour lines indicate the 2-dimensional density of case-level scores within each prompt category. Plots are faceted by model (columns) and dataset (rows). p-values are calculated based on a two-sided multivariate Kruskal-Wallis test at the 0.05 level. Larger is better (more repeatable) for both axes. CoT = chain-of-thought; UDN = Undiagnosed Diseases Network.

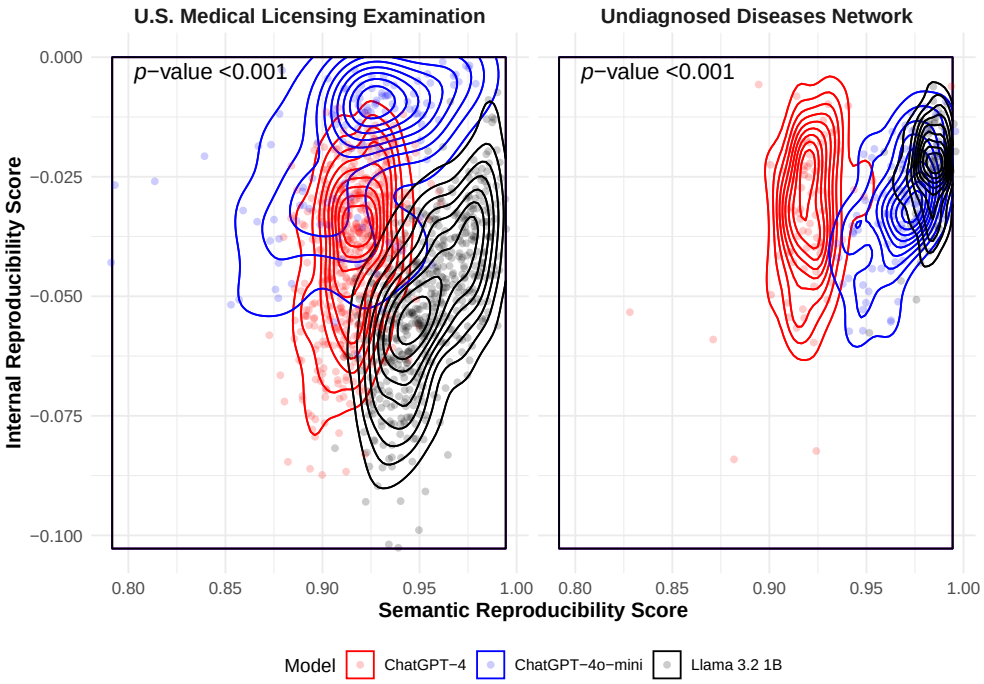

Fig. 2 Semantic and internal reproducibility scores by model and dataset. Each point represents a single clinical case evaluated by a given model, with semantic reproducibility (x-axis) and internal reproducibility (y-axis) computed across the five prompts in Table 1. Contour lines represent 2-dimensional kernel density estimates of case-level reproducibility scores within each model. Left panel = MedQA dataset. Right panel = UDN dataset. p-values are calculated based on a two-sided multivariate Kruskal-Wallis test at the 0.05 level. Larger is better (more reproducible) for both axes. *UDN* = Undiagnosed Diseases Network.

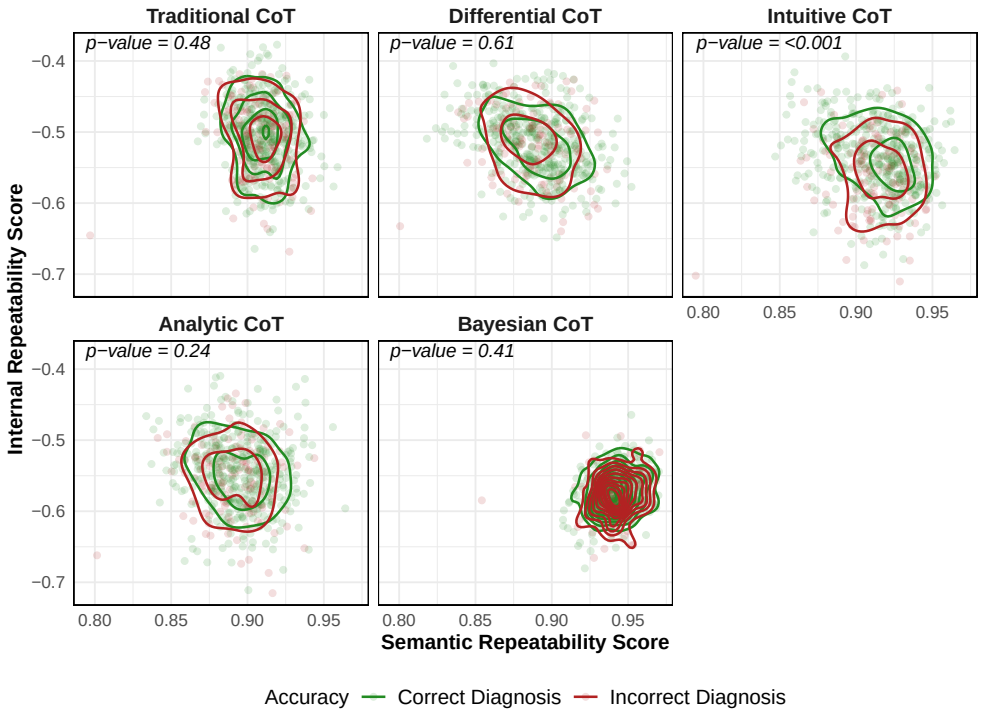

Fig. 3 Semantic and internal repeatability scores for ChatGPT-4 on the MedQA dataset stratified by prompt. Each point represents a single clinical case evaluated by ChatGPT-4, with semantic repeatability score on the x-axis and internal repeatability score on the y-axis. Contour lines indicate the 2-dimensional density of case-level scores for cases that were correctly diagnosed by ChatGPT-4 (green) and incorrectly diagnosed by ChatGPT-4 (red). p-values are calculated based on a two-sided multivariate Kruskal-Wallis test at the 0.05 level. Larger is better (more repeatable) for both axes. CoT = chain-of-thought.

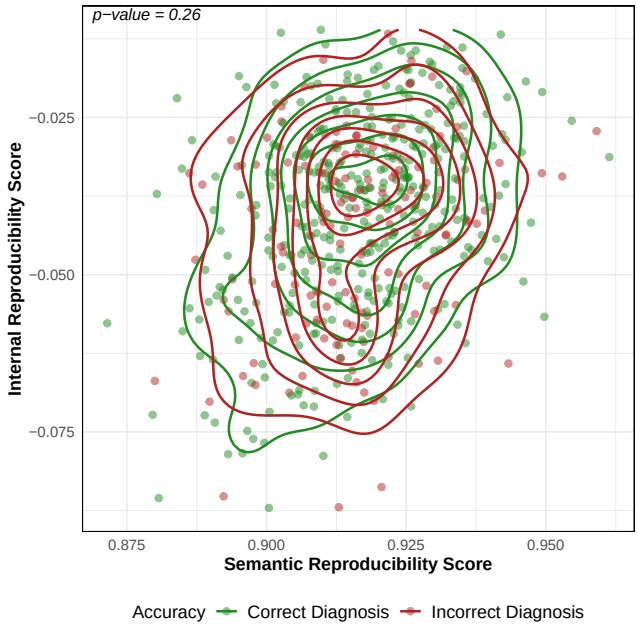

Fig. 4 Semantic and internal reproducibility scores for ChatGPT-4 on the MedQA dataset. Each point represents a single clinical case evaluated by ChatGPT-4, with semantic reproducibility score on the x-axis and internal reproducibility score on the y-axis computed across the five prompts in Table 1. Contour lines indicate the 2-dimensional density of case-level scores for cases that were correctly diagnosed by ChatGPT-4 (green) and incorrectly diagnosed by ChatGPT-4 (red). p-values are calculated based on a two-sided multivariate Kruskal-Wallis test at the 0.05 level. Larger is better (more reproducible) for both axes.

4 Discussion

Our framework focuses on measuring consistency not only as a linguistic property, but also as a dimension of clinical reliability. By evaluating both axes of semantic and internal consistency across models, prompts, and patient cases, our framework enables more clinically meaningful evaluation of LLM behavior in diagnostic tasks. To this end, our framework has direct implications for model selection, regulatory evaluation, and clinical integration of AI tools. Consistency metrics can guide the choice of models and prompting strategies that generate stable, reproducible outputs to support clinical decision-making. Importantly, our framework is model- and prompt-agnostic, making it broadly applicable across different LLMs, reasoning paradigms, and clinical contexts. These metrics also align with the U.S. FDA's draft guidance for AI-enabled medical software, which emphasizes the importance of repeatability and reproducibility in safety and effectiveness assessments [11].

While diagnostic accuracy remains the primary criterion for diagnostic support, it is not sufficient on its own. Once a model demonstrates acceptable accuracy, consistency becomes critical for ensuring that the model's outputs are stable and interpretable over repeated use. Our results show that consistency varies across models, prompts, and the complexity of clinical cases, underscoring that consistency should be evaluated in a context-specific manner. Prompts invoking probabilistic reasoning yielded higher semantic stability for ChatGPT-4, suggesting that consistency is not only model-dependent but also sensitive to how reasoning is elicited. Notably, consistency did not correlate with diagnostic accuracy. This highlights a critical distinction: producing a correct diagnosis is not equivalent to producing that diagnosis reliably. That is, an LLM that generates variable outputs, even if accurate on average, may undermine clinician trust, disrupt decision-making, and limit the model's utility in clinical settings. Therefore, evaluating consistency in conjunction with accuracy is essential for characterizing model behavior and determining its suitability for clinical integration.

This work builds upon recent studies that proposed complementary frameworks to quantify and improve LLM consistency in general natural language generation. For instance, Raj *et al.* [19] proposed semantic consistency metrics and demonstrated that consistency and accuracy are independent properties, an observation that aligns with our findings in the clinical domain. Similarly, Cui *et al.* [20] proposed a divide-and-conquer approach to improve consistency at the sentence level, demonstrating that consistency metrics can serve as both evaluative tools and levers for model refinement.

This work has several limitations that we highlight for future investigation. First, due to the substantial computational burden of generating repeated outputs across prompts, models, and clinical cases, our evaluation prioritized breadth across model types and clinical scenarios rather than exhaustive coverage of all possible configurations. While we selected diverse models, reasoning paradigms, and datasets to reflect real-world clinical variation, additional clinical contexts should be considered. A second potential limitation of this work is that our semantic consistency evaluations used a specific embedding model to quantify output meaning. While results may vary across embedding spaces, all evaluations in this study used the same embedding

model to ensure fair and consistent comparisons. Still, future studies should explore the sensitivity of semantic reproducibility scores to alternative embeddings.

Future extensions of this work includes incorporating recent techniques from hallucination detection and factuality checking, such as sampling-based consistency methods and retrieval-grounded evaluation frameworks, to broaden the scope of consistency evaluation in clinical applications [15, 21–23]

Acknowledgments

The authors are grateful to the patients for participating in the Undiagnosed Diseases Network.

Data availability

The MedQA (U.S. Medical Licensing Examination) cases used in this study are publicly available in Savage *et al.* [3]'s Supplementary Data 1. The Undiagnosed Diseases Network data used in this study contain sensitive patient information. De-identified patient data, including phenotypic and genomic data, are deposited in the database of Genotypes and Phenotypes (dbGaP) maintained by the National Institutes of Health. To explore data available in the latest release, visit the UDN study page in dbGaP. Individuals interested in accessing UDN data through dbGaP should submit a data access request. Detailed instructions for this process can be found on the NIH Scientific Data Sharing website: How to Request and Access Datasets from dbGaP.

Code availability

Code used in this study are publicly available at `https://github.com/cathyshyr/repeatability_and_reproducibility_of_LLMs`.

Competing interests

The authors declare no competing interests.

Funding

This work was supported in part by the National Institutes of Health Common Fund, grant 15-HG-0130 from the National Human Genome Research Institute, U01NS134349 from the National Institute of Neurological Disorders and Stroke, R00LM014429 from the National Library of Medicine, and the Potocsnak Center for Undiagnosed and Rare Disorders.

References

[1] Williams, C.Y., Subramanian, C.R., Ali, S.S., Apolinario, M., Askin, E., Barish, P., Cheng, M., Deardorff, W.J., Donthi, N., Ganeshan, S., et al.: Physician-and large language model–generated hospital discharge summaries. JAMA Internal Medicine (2025)

[2] Small, W.R., Wiesenfeld, B., Brandfield-Harvey, B., Jonassen, Z., Mandal, S., Stevens, E.R., Major, V.J., Lostraglio, E., Szerencsy, A., Jones, S., *et al.*: Large language model–based responses to patients' in-basket messages. JAMA network open **7**(7), 2422399–2422399 (2024)

[3] Savage, T., Nayak, A., Gallo, R., Rangan, E., Chen, J.H.: Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. NPJ Digital Medicine **7**(1), 20 (2024)

[4] Kung, T.H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., *et al.*: Performance of chatgpt on usmle: potential for ai-assisted medical education using large language models. PLoS digital health **2**(2), 0000198 (2023)

[5] Strong, E., DiGiammarino, A., Weng, Y., Kumar, A., Hosamani, P., Hom, J., Chen, J.H.: Chatbot vs medical student performance on free-response clinical reasoning examinations. JAMA internal medicine **183**(9), 1028–1030 (2023)

[6] Kanjee, Z., Crowe, B., Rodman, A.: Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. Jama **330**(1), 78–80 (2023)

[7] Barile, J., Margolis, A., Cason, G., Kim, R., Kalash, S., Tchaconas, A., Milanaik, R.: Diagnostic accuracy of a large language model in pediatric case studies. JAMA pediatrics **178**(3), 313–315 (2024)

[8] Elazar, Y., Kassner, N., Ravfogel, S., Ravichander, A., Hovy, E., Schütze, H., Goldberg, Y.: Measuring and improving consistency in pretrained language models. Transactions of the Association for Computational Linguistics **9**, 1012–1031 (2021)

[9] Gu, B., Desai, R.J., Lin, K.J., Yang, J.: Probabilistic medical predictions of large language models. npj Digital Medicine **7**(1), 367 (2024)

[10] Ceballos-Arroyo, A.M., Munnangi, M., Sun, J., Zhang, K., Mcinerney, J., Wallace, B.C., Amir, S.: Open (clinical) llms are sensitive to instruction phrasings. In: Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, pp. 50–71 (2024)

[11] U.S. Food and Drug Administration: Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission

Recommendations. Draft Guidance for Industry and Food and Drug Administration Staff. Contains nonbinding recommendations. Draft—Not for implementation. Distributed for comment purposes only. (2025). https://www.fda.gov/regulatory-information/search-fda-guidance-documents/artificial-intelligence-enabled-device-software-functions-lifecycle-management-and-marketing

[12] Perlis, R.H., Fihn, S.D.: Evaluating the application of large language models in clinical research contexts. JAMA Network Open **6**(10), 2335924–2335924 (2023)

[13] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM computing surveys **55**(12), 1–38 (2023)

[14] Manakul, P., Liusie, A., Gales, M.J.: Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. arXiv preprint arXiv:2303.08896 (2023)

[15] Zhang, J., Li, Z., Das, K., Malin, B.A., Kumar, S.: Sac3: reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. arXiv preprint arXiv:2311.01740 (2023)

[16] Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., Szolovits, P.: What disease does this patient have? a large-scale open domain question answering dataset from medical exams. Applied Sciences **11**(14), 6421 (2021)

[17] Ramoni, R.B., Mulvihill, J.J., Adams, D.R., Allard, P., Ashley, E.A., Bernstein, J.A., Gahl, W.A., Hamid, R., Loscalzo, J., McCray, A.T., *et al.*: The undiagnosed diseases network: accelerating discovery about health and disease. The American Journal of Human Genetics **100**(2), 185–192 (2017)

[18] Balachandran, A.: MedEmbed: Medical-Focused Embedding Models. https://github.com/abhinand5/MedEmbed

[19] Raj, H., Gupta, V., Rosati, D., Majumdar, S.: Semantic consistency for assuring reliability of large language models. arXiv preprint arXiv:2308.09138 (2023)

[20] Cui, W., Li, Z., Lopez, D., Das, K., Malin, B., Kumar, S., Zhang, J.: Divide-conquer-reasoning for consistency evaluation and automatic improvement of large language models. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, pp. 334–361 (2024)

[21] Wang, X., Yan, Y., Huang, L., Zheng, X., Huang, X.-J.: Hallucination detection for generative large language models by bayesian sequential estimation. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 15361–15371 (2023)

[22] Augenstein, I., Baldwin, T., Cha, M., Chakraborty, T., Ciampaglia, G.L., Corney, D., DiResta, R., Ferrara, E., Hale, S., Halevy, A., *et al.*: Factuality challenges in the era of large language models and opportunities for fact-checking. Nature Machine Intelligence **6**(8), 852–863 (2024)

[23] Lee, D., Yu, H.: Refind at semeval-2025 task 3: Retrieval-augmented factuality hallucination detection in large language models. arXiv preprint arXiv:2502.13622 (2025)