A statistical framework for evaluating the repeatability and reproducibility of large language models

Cathy Shyr[1,2,3*], Boyu Ren[4], Chih-Yuan Hsu[2], Chao Yan[1], Rory J. Tinker[5], Thomas A. Cassini[3], Rizwan Hamid[3], Adam Wright[1,6], Lisa Bastarache[1], Josh F. Peterson[1,6], Bradley A. Malin[1,2,7], Hua Xu[8]

[1*]Department of Biomedical Informatics, Vanderbilt University Medical Center, 2525 West End Avenue, Nashville, 37203, TN, USA.
[2]Department of Biostatistics, Vanderbilt University Medical Center, 2525 West End Avenue, Nashville, 37203, TN, USA.
[3]Department of Pediatrics, Vanderbilt University Medical Center, 2200 Children's Way, Nashville, 37232, TN, USA.
[4]Laboratory for Psychiatric Biostatistics, McLean Hospital, 115 Mill Street, Belmont, 02478, MA, USA.
[5]Department of Medical Genetics and Genomics, Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Pl, New York, 10028, New York, USA.
[6]Department of Medicine, Vanderbilt University Medical Center, 1161 21st Ave S, Nashville, 37232, TN, USA.
[7]Department of Computer Science, Vanderbilt University, 1400 18th Avenue S, Nashville, 37212, TN, USA.
[8]Department of Biomedical Informatics and Data Science, Yale School of Medicine, 101 College Street, New Haven, 06520, CT, USA.

*Corresponding author(s). E-mail(s): cathy.shyr@vumc.org;
Contributing authors: bor158@mail.harvard.edu; chih-yuan.hsu@vumc.org; chao.yan.1@vumc.org; rory.tinker@mssm.edu; thomas.a.cassini@vumc.org; rizwan.hamid@vumc.org; adam.wright@vumc.org; lisa.bastarache@vumc.org; josh.peterson@vumc.org; b.malin@vumc.org; hua.xu@yale.edu;

Abstract

Objective. Systematic evaluation of variability in large language model (LLM)-generated outputs is critical for assessing their reliability. We developed a regulatory-informed statistical framework to quantify LLM repeatability and reproducibility.

Materials and Methods. Using definitions from U.S. Food and Drug Administration draft guidance for AI-enabled medical software, our framework quantifies LLM repeatability and reproducibility along two complementary dimensions: semantic and internal. Semantic metrics quantify variability in the meaning of LLM-generated outputs, and internal metrics quantify variability in token-level probability distributions during text generation. We applied the framework to diagnostic reasoning questions across multiple LLMs and prompting strategies using two datasets: 518 U.S. Medical Licensing Examination (USMLE) questions and 90 real-world rare disease cases from the Undiagnosed Diseases Network (UDN).

Results. LLM repeatability and reproducibility varied by dataset, prompting strategy, and model. Prompts eliciting Bayesian diagnostic reasoning achieved significantly higher semantic repeatability than other prompting strategies ($\boldsymbol{p < 0.001}$). LLM repeatability and reproducibility were generally not associated with diagnostic accuracy.

Discussion. LLM repeatability and reproducibility depend not only on the model, but also on the prompting strategy used to elicit reasoning. A model may produce a correct response in a single run but not consistently reproduce that response across repeated runs.

Conclusion. We developed a general framework for quantifying LLM repeatability and reproducibility across models, prompting strategies, and biomedical tasks. The metrics in this framework can be used to systematically compare output variability across different LLMs, prompts, model configurations, or evaluation datasets.

Keywords: repeatability, reproducibility, large language model, artificial intelligence, diagnostic reasoning

1 INTRODUCTION

Large language models (LLMs) are increasingly applied in biomedical settings and have demonstrated promising performance across clinical applications, including generating clinical documentation, responding to patient messages, automating phenotyping, and supporting clinical decision-making [1–13]. Evaluation of LLMs has largely focused on task-level metrics such as accuracy. However, accuracy alone does not capture how consistently a model behaves across repeated runs. Because LLMs generate text by sampling tokens from probability distributions, identical prompts can yield different outputs across runs [14]. Therefore, a model may produce a correct response in a single run but fail to reproduce that response consistently across repeated runs, potentially complicating interpretation of model outputs in clinical settings [15]. Despite the

distinction between accuracy and output variability, systematic approaches for quantifying variability of LLM outputs remain limited, leaving an important gap in the comprehensive evaluation of model performance.

The importance of measuring variability of LLM outputs has been increasingly emphasized by regulatory agencies and research communities. Draft guidance from the U.S. Food and Drug Administration (FDA) on AI-enabled medical software recommends quantifying model variability based on repeatability and reproducibility [16]. Medical journals recommend researchers to assess variability as a critical component when evaluating the application of LLMs in clinical research [17]. In addition, reporting standards, such as CONSORT-AI and DECIDE-AI, emphasize reproducibility and rigorous evaluation of AI systems in clinical studies [18, 19]. Collectively, these recommendations underscore the importance of systematically assessing variability as part of the comprehensive evaluation of LLM performance.

Variability is common in biomedical settings, as different clinicians may document cases differently or arrive at different diagnoses depending on their training or specialty [20]. The key distinction is that clinicians can explain their reasoning and contextualize such differences, whereas LLMs lack transparent indicators of how much their responses vary across repeated runs [21]. For example, an LLM used for diagnostic support may correctly identify a patient's diagnosis in one run but generate a different set of candidate diagnoses when given the same patient case again. Both outputs may appear plausible, yet their inconsistency can complicate interpretation and reduce confidence in the model's recommendations. Without systematic methods to quantify this variability, it is difficult for users to assess the consistency of model outputs across repeated use. Quantifying variability is therefore critical for evaluating the robustness of LLMs and complements accuracy-based metrics to provide a more comprehensive assessment of model performance.

Existing metrics commonly used for evaluating LLM-generated outputs, including BLEU, ROUGE, and BERTScore, were developed for natural language generation tasks such as machine translation and summarization [22–24]. These metrics assess similarity between a model's output and a reference text based on lexical overlap or semantic similarity. While such metrics are useful for evaluating the accuracy or relevance of model-generated text relative to a ground truth reference, they were not designed to quantify variability across repeated runs. In parallel, prompt engineering has received increasing attention, with studies proposing different strategies such as chain-of-thought prompting for complex reasoning, few-shot prompting for task adaptation, and self-consistency prompting for improving reasoning accuracy [25–29]. However, these approaches typically evaluate prompts based on accuracy-related performance metrics rather than run-to-run variability. As a result, existing methods provide limited insight into how consistently an LLM produces outputs across multiple runs, which is critical for assessing model robustness and reliability in biomedical applications.

To address this gap, our objective was to develop a regulatory-informed statistical framework for quantifying variability of LLM outputs. Consistent with FDA recommendations for evaluating AI-enabled software, our framework operationalizes the

evaluation of repeatability, defined as the agreement of model outputs under identical conditions, and reproducibility, agreement under pre-specified, different conditions (e.g., different users or experimental setup) [16]. Within these concepts, our framework further defines two complementary dimensions of variability, semantic and internal. These dimensions yield a total of four metrics: 1) Semantic Repeatability, 2) Internal Repeatability, 3) Semantic Reproducibility, and 4) Internal Reproducibility. Semantic metrics quantify variability in the meaning of LLM outputs across runs, whereas internal metrics quantify variability in the model's token-level probability distributions, which reflects the stability of the model's text-generation process. Importantly, our framework is agnostic to both the prompt and the specific model within the class of autoregressive LLMs, which include most modern systems such as ChatGPT, Claude, LLaMA, Qwen, and Gemini [30–32]. We demonstrate the framework using diagnostic reasoning as a use case, evaluating LLM outputs on U.S. Medical Licensing Examination questions and real-world rare disease patient cases from the Undiagnosed Diseases Network. More broadly, the proposed framework provides a systematic approach for quantifying variability of LLM outputs, enabling more comprehensive evaluation of model performance in biomedical research and applications.

2 MATERIALS AND METHODS

2.1 Definitions of Repeatability and Reproducibility

Motivated by the U.S. FDA's draft guidance on AI-enabled medical software, which recommends evaluation of both repeatability and reproducibility in AI systems, our framework defines and operationalizes four metrics: 1) Semantic Repeatability, 2) Internal Repeatability, 3) Semantic Reproducibility, and 4) Internal Reproducibility (**Figure 1**). Repeatability is defined by the FDA as "the closeness of agreement of repeated measurements taken under the same conditions." [16] In our framework, repeatability refers to generating LLM outputs across repeated runs using the same model, prompt, and generation parameters, and measuring the agreement across these outputs. Reproducibility, in contrast, is defined by the FDA as "the closeness of agreement of repeated measurements taken under different, pre-specified conditions." [16] In our framework, reproducibility refers to generating outputs under controlled variation in one or more experimental conditions while holding other factors constant. These conditions may include, for example, different prompts, users, or care settings.

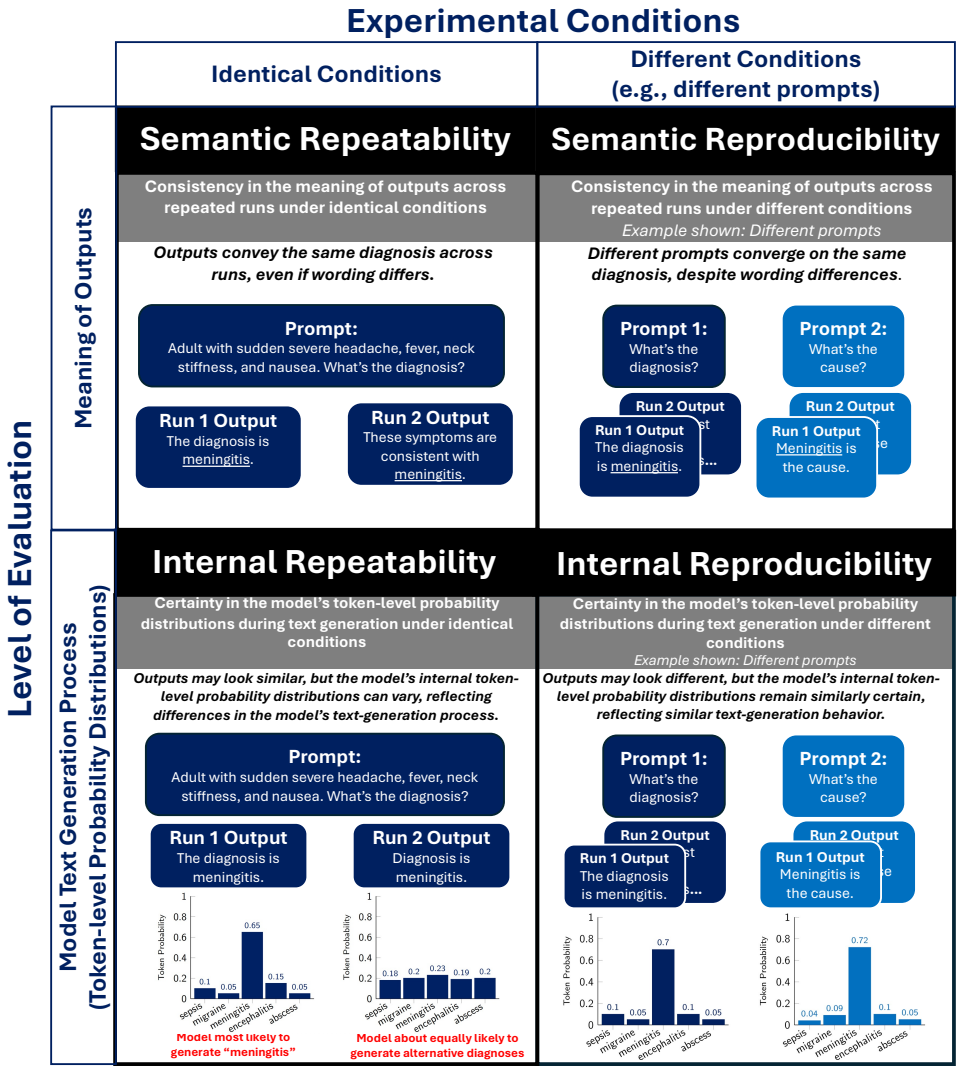

Fig. 1: **Overview of LLM variability metrics.** (1) Semantic Repeatability: consistency of meaning across repeated runs under identical conditions. (2) Semantic Reproducibility: consistency of meaning across repeated runs under different, pre-specified conditions (illustrated here using different prompts as an example). (3) Internal Repeatability: certainty in the token-level probability distributions during text generation under identical conditions. (4) Internal Reproducibility: certainty in the token-level probability distributions during text generation under different, pre-specified conditions (illustrated here using different prompts as an example).

Within both repeatability and reproducibility, we evaluate LLM outputs along two complementary dimensions: semantic and internal metrics (**Figure 1**). Semantic metrics quantify variability in the meaning of outputs across repeated runs, capturing a clinically relevant property that assesses meaning rather than surface-level wording differences. Internal metrics, by contrast, quantify variability in the token-level probability distributions that underlie LLM text generation. During text generation, autoregressive LLMs (e.g., ChatGPT, LLaMA, etc.) sample the next token from a probability distribution conditioned on prior context. Consequently, two outputs may appear nearly identical on the surface (e.g., "The diagnosis is meningitis" and "Diagnosis is meningitis") yet differ in their underlying token-level probability distributions during text generation. For example, one run may assign a high probability to the diagnosis "meningitis" while another assigns roughly equal probabilities across several candidate diagnoses (**Figure 1**). Such differences reflect variability in the LLM's internal text-generation process.

2.2 Statistical Framework for Evaluating LLM Repeatability and Reproducibility

In this section, we formally define the four proposed metrics of LLM variability. Let X denote the input prompt and $\mathcal{V}$ the LLM's output vocabulary (i.e., the set of all possible output tokens). Let $Y_{r,i} \in \mathcal{V}$ denote the output token generated by the LLM at position $i = 1 \ldots, L_r$, where L_r is the length of the output sequence in run $r = 1 \ldots,\ R$, and R is the total number of runs. We denote the full output sequence generated by an LLM in run r as $Y_{r,1:L_r}$.

2.2.1 Semantic Repeatability Score

The Semantic Repeatability Score quantifies the consistency in the meaning of LLM outputs across repeated runs under identical conditions (e.g., same prompt, LLM, text-generation parameters, etc.). Formally, let $\mathcal{V}^*$ denote the set of all finite-length token sequences over the vocabulary $\mathcal{V}$. Let $\mathcal{E} : \mathcal{V}^* \to \mathbb{R}$ be an embedding function that maps an output sequence to a d-dimensional vector representation. The embedding function may be implemented using any suitable text embedding model and is not required to be the same model that generated the outputs. For each run $r = 1, \ldots, R$, we compute the embedding vector $e_r = \mathcal{E}\left(Y_{r,1:L_r}\right) \in \mathbb{R}$. We define the Semantic Repeatability Score in **Definition 1**.

Definition 1. Semantic Repeatability Score. The Semantic Repeatability Score quantifies the consistency in the meaning of LLM outputs across repeated runs under identical conditions. Formally, it's defined as the average pairwise cosine similarity between the embedding representations of LLM outputs across runs:

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

For interpretability, we rescale $\bar{S}^{\mathrm{Rpt}}$ to the interval $[0, 1]$:

$$\widetilde{S}^{\mathrm{Rpt}} = \frac{\bar{S}^{\mathrm{Rpt}} + 1}{2}, \qquad 0 \leq \widetilde{S}^{\mathrm{Rpt}} \leq 1.$$

Larger values of $\widetilde{S}^{\mathrm{Rpt}}$ indicate greater semantic consistency across LLM outputs (i.e., larger = more repeatable).

2.2.2 Internal Repeatability Score

The Internal Repeatability Score quantifies the certainty of an autoregressive LLM's token-level probability distributions during text generation under identical conditions (e.g., same prompt, LLM, text-generation parameters, etc.). Unlike the Semantic Repeatability Score, which evaluates the semantic similarity of final outputs, the Internal Repeatability Score characterizes the internal probabilistic sampling mechanisms used by autoregressive LLMs during text generation.

To define the Internal Repeatability Score, we first outline the text-generation process of autoregressive LLMs. Autoregressive LLMs, which include most modern systems such as ChatGPT, Gemini, Qwen, and LLaMA, generate text sequentially by sampling tokens one at a time [30–32]. At each step, the model produces a probability distribution over its vocabulary and samples the next token from that distribution, conditioned on the prompt and previously generated tokens. Formally, this process can be described as follows:

Step 1. (User provides a prompt): The user provides an input prompt X.

Step 2. (Logit computation): At position i of the generated output sequence in run r, the LLM produces a vector of logits over the vocabulary $\mathcal{V}$:

$$z_{r,i} = (z_{r,i}(v) : v \in \mathcal{V}) \in \mathbb{R}^{|\mathcal{V}|}.$$

These logits are produced by the LLM decoding function

$$f_i : (X, Y_{r,1:i-1}) \mapsto z_{r,i},$$

where $Y_{r,1:i-1}$ denotes the history of tokens generated up to position $i - 1$. At the first position ($i = 1$), the logits depend only on the user's prompt:

$$z_{r,1} = f_1(X).$$

At subsequent positions ($i = 2, \ldots, L_r$), the logits depend on both the prompt and previously generated tokens:

$$z_{r,i} = f_i(X, Y_{r,1:i-1}), \qquad i = 2, \ldots, L_r.$$

Step 3. (Conversion to token-level probability distributions): Let T denote the user-specified temperature parameter in LLMs that controls the randomness of token sampling. The logits are transformed into a probability

distribution over the vocabulary using the temperature-scaled softmax function:

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

Larger values of T produce flatter probability distributions over all possible tokens (i.e., more uncertainty), whereas smaller values produce more peaked distributions (i.e., specific tokens have a higher probability of being sampled).

Step 4. (Top-k truncation): To focus on the tokens most likely to be sampled and reduce the influence of extremely low-probability tokens, users can specify the parameter, top-k, to truncate the probability distribution $\pi_{r,i}(\cdot)$ to the k tokens with the highest probabilities of being sampled. Let $K_{r,i} \subset \mathcal{V}$ denote this set of tokens. The top-k truncated probability distribution is

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

Step 5. (Token sampling): The next output token is sampled from the truncated probability distribution $\widetilde{\pi}_{r,i}(\cdot)$.

Steps 2 through 5 repeats until the LLM generates the last token in the output sequence. To quantify the uncertainty of these token-level probability distributions, we calculate the Shannon entropy of the truncated distribution at each position:

$$H_{r,i} = -\sum_{v \in K_{r,i}} \widetilde{\pi}_{r,i}(v) \log \widetilde{\pi}_{r,i}(v), \qquad i = 1, \ldots, L_r.$$

Lower entropy corresponds to more peaked (more certain) token-level probability distributions, whereas higher entropy indicates greater uncertainty across candidate tokens. We average entropy across output positions to obtain the run-level entropy,

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

We define the Internal Repeatability Score in **Definition 2**

Definition 2. Internal Repeatability Score. The Internal Repeatability Score quantifies the certainty of an LLM's token-level probability distributions during text generation under identical conditions (e.g., same prompt, model, and generation parameters). To define it, we first compute the mean entropy across R independent runs as

$$\bar{H} = \frac{1}{R} \sum_{r=1}^{R} H_r,$$

where H_r is the run-level entropy in run r. Because the entropy of a distribution supported on at most k tokens is bounded above by $\log_2 k$, the mean entropy satisfies $0 \leq \bar{H} \leq \log_2 k$. We then rescale $\bar{H}$ to the interval $[0, 1]$ to obtain the Internal Repeatability Score,

$$\widetilde{H}^{\mathrm{Rpt}} = 1 - \frac{\bar{H}}{\log_2 k}, \qquad 0 \leq \widetilde{H}^{\mathrm{Rpt}} \leq 1.$$

Larger values of $\widetilde{H}^{\mathrm{Rpt}}$ indicate greater certainty in the model's token-level probability distributions during text generation under identical conditions (i.e., larger = more repeatable).

2.2.3 Semantic Reproducibility Score

The Semantic Reproducibility Score quantifies the consistency in the meaning of LLM outputs across repeated runs under different, pre-specified experimental conditions. These conditions may include, for example, different prompts, users, LLMs, or care settings. Let P denote the number of experimental conditions considered (e.g., P different prompts). For each condition $p = 1, \ldots, P$, we generate R independent runs and compute the embedding of each generated output using the embedding function $\mathcal{E}$. Formally, for run r under condition p, we compute the embedding

$$e_r^{(p)} = \mathcal{E}\left(Y_{r,1:L_r}^{(p)}\right) \in \mathbb{R}.$$

We then compute the average embedding across runs for each condition p,

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

Using these condition-level embeddings, we define the Semantic Reproducibility Score in **Definition 3**.

Definition 3. Semantic Reproducibility Score. The Semantic Reproducibility Score quantifies the consistency in the meaning of LLM outputs across repeated runs under different, pre-specified experimental conditions. Formally, it's defined as the average pairwise cosine similarity between the mean embedding representations of outputs across experimental conditions:

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

For interpretability, we rescale $\bar{S}^{\mathrm{Rpd}}$ to the interval $[0, 1]$:

$$\widetilde{S}^{\mathrm{Rpd}} = \frac{\bar{S}^{\mathrm{Rpd}} + 1}{2}, \qquad 0 \leq \widetilde{S}^{\mathrm{Rpd}} \leq 1$$

Larger values of $\widetilde{S}^{\mathrm{Rpd}}$ indicate greater consistency in the semantic meaning of LLM outputs across experimental conditions (i.e., larger = more reproducible).

2.2.4 Internal Reproducibility Score

The Internal Reproducibility Score quantifies the consistency in the certainty of an autoregressive LLM's token-level probability distributions during text generation under different, pre-specified experimental conditions. Let P denote the number of experimental conditions considered (e.g., P different prompts). For each condition $p = 1, \ldots, P$, we generate R independent runs and compute the run-level entropy $H_r^{(p)}$ using the same procedure defined for the Internal Repeatability Score. Here, $H_r^{(p)}$ denotes the mean token-level entropy of the truncated probability distributions generated during run r under condition p. We then compute the average run-level entropy across runs for each condition,

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

Using these condition-level entropy estimates, we define the Internal Reproducibility Score in **Definition 4**.

Definition 4. Internal Reproducibility Score. The Internal Reproducibility Score quantifies the certainty of token-level probability distributions during text generation under different, pre-specified experimental conditions. Formally, it is defined as the average pairwise difference in mean entropy between conditions, rescaled to the interval $[0, 1]$ for interpretability:

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

$$\widetilde{H}^{\mathrm{Rpd}} = 1 - \frac{\bar{H}^{\mathrm{Rpd}}}{\log_2 k}, \qquad 0 \le \widetilde{H}^{\mathrm{Rpd}} \le 1.$$

where $\bar{H}^{(p)}$ denotes the mean run-level entropy under experimental condition p. Larger values of $\widetilde{H}^{\mathrm{Rpd}}$ indicate greater consistency in the certainty of the model's token-level probability distributions across experimental conditions (i.e., larger = more reproducible).

2.3 Empirical Evaluation of LLM Diagnostic Reasoning

To demonstrate the practical application of the proposed repeatability and reproducibility metrics, we performed an empirical evaluation of LLM diagnostic reasoning across multiple clinical datasets, prompting strategies, and model architectures. In this empirical use case, reproducibility was operationalized using different diagnostic reasoning prompts. The proposed framework, however, is general and can be applied

to quantify reproducibility across any pre-specified experimental condition, such as different LLMs, users, or care settings.

2.3.1 Diagnostic Reasoning Prompts

We evaluated LLM repeatability and reproducibility across five chain-of-thought (CoT) diagnostic reasoning prompts developed and validated by Savage *et al* [33]. These prompts were designed to elicit distinct diagnostic reasoning strategies used in clinical practice, including intuitive reasoning, analytic reasoning, and probabilistic reasoning (**Table 1**). The full prompts are provided in the **Supplementary Materials**.

Table 1: **Overview of empirical evaluation.** CoT = chain of thought; USMLE = U.S. Medical Licensing Examination; NIH = National Institutes of Health; LLM = Large language model

Prompt Name	**Prompt from Savage *et al.*[33]**	**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[34] 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 cases
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 [33]
ChatGPT-4o-mini	Commercial	Lightweight and efficient
Llama 3.2-1B	Open Source	Lightweight and efficient

2.3.2 Datasets

To evaluate repeatability and reproducibility across diverse clinical contexts, we selected two complementary datasets: 1) a standardized benchmark for medical knowledge (MedQA), and 2) real-world rare disease cases from the Undiagnosed Diseases Network (UDN) [34, 35].

MedQA is a publicly available benchmark consisting of diagnostic clinical vignettes derived from U.S. Medical Licensing Examination (USMLE) questions [34]. Following Savage *et al.*, we used the same set of 518 questions, reformulated from multiple-choice format into free-response prompts [33]. We focused on Step 2 and Step 3 cases, which emphasize clinical reasoning over rote recall. These vignettes are fully specified and standardized, making them well-suited for controlled benchmarking of diagnostic reasoning. However, they may be idealized relative to real-world clinical scenarios. An example is shown in **Table 2**. All questions are publicly available and provided in **Supplementary Data 1** of Savage *et al* [33].

To complement MedQA and address concerns that public benchmarks may have appeared during LLM pre-training, we additionally analyzed 90 non-public rare disease cases from the Vanderbilt University Medical Center (VUMC) site of the UDN [36]. In contrast to the exam-style vignettes of USMLE, UDN patients often present with complex and heterogeneous phenotypes, frequently involving multi-system manifestations, non-diagnostic clinical test results, and atypical genotypes that are challenging to interpret (e.g., variants of uncertain significance) [37, 38]. They also may have symptoms or test results included in the case data that are unrelated to their final diagnosis. Each UDN case is summarized as a multi-paragraph narrative detailing the patient's medical history and prior workup. An illustrative example is shown in **Table 2**. This study was approved by the VUMC Institutional Review Board (IRB# 172005).

2.3.3 LLMs

We evaluated three LLMs selected to represent a range of model sizes (**Table 1**): 1) ChatGPT-4 (OpenAI) is a commercial LLM that has previously demonstrated strong diagnostic reasoning performance on the same USMLE dataset used in this study [33]; 2) ChatGPT-4o-mini (OpenAI) is a smaller, more cost-efficient variant of GPT-4 designed for more lightweight deployments. Including this model allows us to evaluate repeatability and reproducibility under a lightweight, resource-efficient setting; 3) LLaMA 3.2-1B (Meta AI) is an open-source, lightweight model that provides a computationally efficient alternative to commercial systems [30].

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

USMLE Case: A man in his fifties 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.
UDN Case: **"One-liner":** A teenage boy 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 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 **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)

2.3.4 Evaluation Setup and Parameter Settings

ChatGPT models were accessed via the Microsoft Azure OpenAI application programming interface, and LLaMA 3.2-1B was deployed on a secure VUMC server. To ensure patient privacy, all evaluations involving UDN cases were conducted using institutionally approved secure computing environments. For all runs, we set the temperature T to 0.5 and top-k to 30. This parameter configuration was chosen to balance determinism and diversity so that both repeatability and reproducibility could be meaningfully evaluated. Extremely deterministic settings could artificially inflate repeatability, whereas highly stochastic settings could exaggerate variability. Although fixed in our evaluation, the proposed framework is agnostic to parameter choice and can be applied under any parameter configuration appropriate to the user's application.

2.3.5 Statistical Analysis

We used two-sided multivariate Kruskal-Wallis tests at the 0.05 level to assess differences in repeatability and reproducibility scores across diagnostic reasoning prompts, datasets (USMLE vs. UDN), and LLMs. When the result from a Kruskal-Wallis test was statistically significant, we performed pairwise post-hoc comparisons using Dunn's test. To account for multiple testing, p-values were adjusted using the Holm-Bonferroni method. Adjusted p-values less than 0.05 were considered statistically significant.

3 RESULTS

3.1 Results from the Empirical Evaluation of LLM Diagnostic Reasoning

Across the five diagnostic reasoning prompts, we evaluated 518 USMLE cases and 90 UDN cases using three LLMs with $R = 100$ independent runs per prompt-case-LLM combination, resulting in $912,000$ total generations. Repeatability metrics (Semantic Repeatability and Internal Repeatability) were computed within each prompt-case-LLM combination across repeated runs. Reproducibility metrics (Semantic Reproducibility and Internal Reproducibility) were computed across prompts for each case-LLM combination.

3.1.1 Repeatability of LLM Diagnostic Reasoning

Overall, semantic repeatability varied across diagnostic reasoning prompts, models, and datasets (**Figure 2**). For ChatGPT-4, the Bayesian CoT prompt tended to produce higher Semantic Repeatability Scores across both USMLE and UDN cases. Differences across prompts were statistically significant for ChatGPT-4o-mini on USMLE cases ($p < 0.001$). In contrast, internal repeatability showed relatively little variation across models and prompting strategies. One exception was ChatGPT-4o-mini, where Traditional CoT and Bayesian CoT prompts had statistically significantly lower Internal Repeatability Scores compared with other prompting strategies (adjusted $p < 0.001$). Across all models, repeatability scores for UDN

cases were more tightly clustered across prompting strategies than for USMLE cases (**Figure 2**), indicating less variation in repeatability across prompts. We provide all repeatability scores in **Table S1**.

3.1.2 Reproducibility of LLM Diagnostic Reasoning

In this empirical evaluation, reproducibility was assessed across different diagnostic reasoning prompts (**Table 1**). Across all three LLMs, reproducibility scores varied less for UDN than for USMLE cases (**Figure 3**). On the USMLE cases, ChatGPT-4o-mini achieved higher Internal Reproducibility Scores compared to the other LLMs (adjusted $p < 0.001$). In contrast, LLaMA 3.2-1B exhibited higher Semantic Reproducibility Scores (adjusted $p < 0.001$). For the UDN cases, Internal Reproducibility Scores were relatively uniform across all three LLMs. However, LLaMA 3.2-1B exhibited higher Semantic Reproducibility scores than the other models (adjusted $p < 0.001$). We provide all reproducibility scores in **Table S1**.

3.1.3 Relationship Among LLM Diagnostic Accuracy, Repeatability, and Reproducibility

We further evaluated whether repeatability and reproducibility were associated with diagnostic accuracy using published ground-truth labels for ChatGPT-4 on USMLE cases from Savage *et al* [33]. In that study, physicians manually reviewed ChatGPT-4's outputs and assigned diagnostic accuracy labels (correctly diagnosed vs. incorrectly diagnosed) for each case (see **Supplementary Information** in Savage *et al.* [33]). Using the same set of USMLE cases and prompts, we generated $R = 100$ outputs per case and computed repeatability and reproducibility scores. We then compared these scores between correctly and incorrectly diagnosed cases. As shown in **Figure 4**, Semantic Repeatability and Internal Repeatability Scores were similar between correctly and incorrectly diagnosed cases across most prompting strategies. For the Traditional CoT, Differential Diagnosis CoT, Analytic CoT, and Bayesian CoT prompts, no statistically significant differences in repeatability scores were observed between the two groups. However, under the Intuitive CoT prompting strategy, correctly diagnosed cases exhibited significantly higher Internal Repeatability Scores than incorrectly diagnosed cases ($p < 0.001$). **Figure 5** shows the corresponding analysis for reproducibility scores. No statistically significant differences in either semantic reproducibility or internal reproducibility were observed between correctly and incorrectly diagnosed cases ($p = 0.26$).

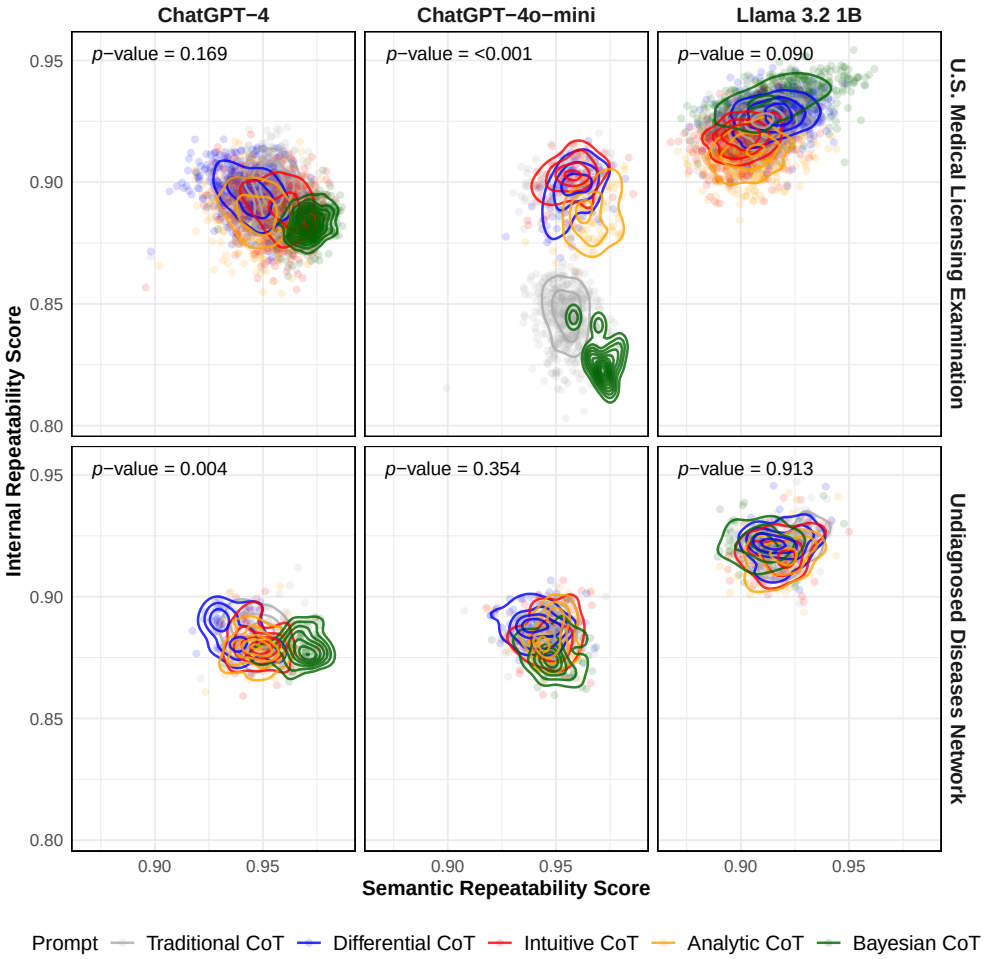

Fig. 2: **Semantic and Internal Repeatability Scores by large language model, dataset, and prompt.** Each dot represents a single clinical case evaluated under a specific prompting strategy, plotted by its Semantic Repeatability Score (x-axis) and Internal Repeatability Score (y-axis). Contour lines represent two-dimensional density estimates for each prompting strategy. Panels are stratified by model (columns) and dataset (rows). P-values are calculated based on a two-sided multivariate Kruskal–Wallis test at the 0.05 level. Higher values indicate greater repeatability for both metrics. *CoT = chain-of-thought; UDN = Undiagnosed Diseases Network.*

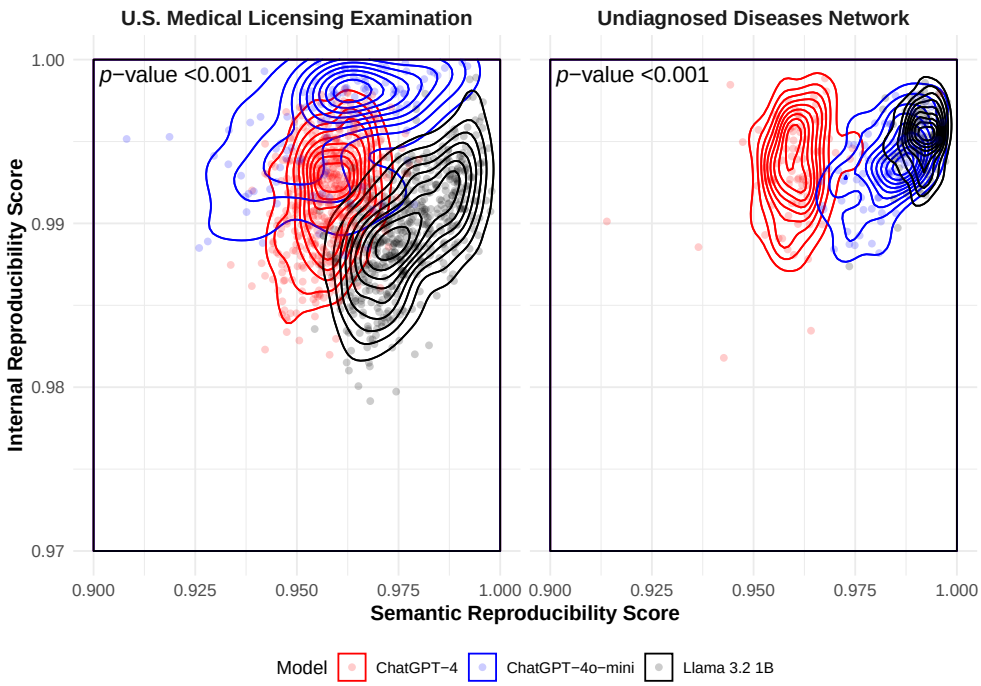

Fig. 3: **Semantic and Internal Reproducibility Scores by large language model and dataset.** Each dot represents a single clinical case evaluated by a model, with Semantic Reproducibility Score (x-axis) and Internal Reproducibility Score (y-axis) computed across the five diagnostic reasoning prompts. Contour lines represent two-dimensional density estimates for each prompting strategy. P-values are calculated based on a two-sided multivariate Kruskal–Wallis test at the 0.05 level. Higher values indicate greater reproducibility for both metrics.

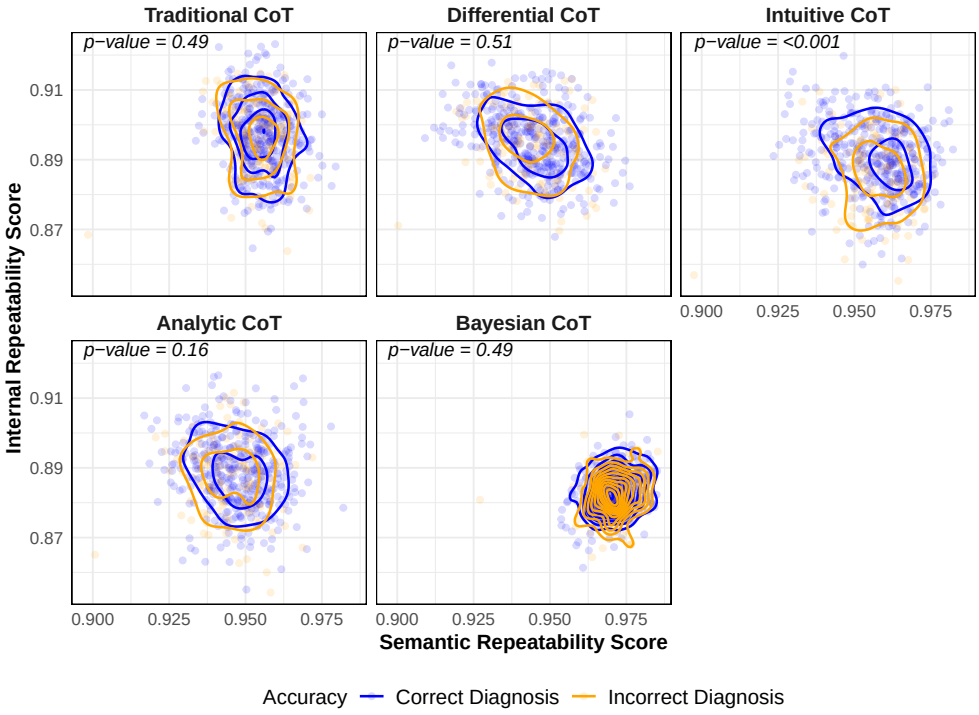

Fig. 4: Semantic and Internal Repeatability Scores for ChatGPT-4 on the MedQA (U.S. Medical Licensing Examination) dataset stratified by prompt. Each dot represents a single clinical case evaluated by ChatGPT-4, with Semantic Repeatability Score on the x-axis and Internal Repeatability Score on the y-axis. Contour lines represent two-dimensional density estimates for cases that were correctly diagnosed by ChatGPT-4 (blue) and incorrectly diagnosed by ChatGPT-4 (orange). P-values are calculated based on a two-sided multivariate Kruskal-Wallis test at the 0.05 level. Higher values indicate greater repeatability for both metrics. *CoT = chain-of-thought.*

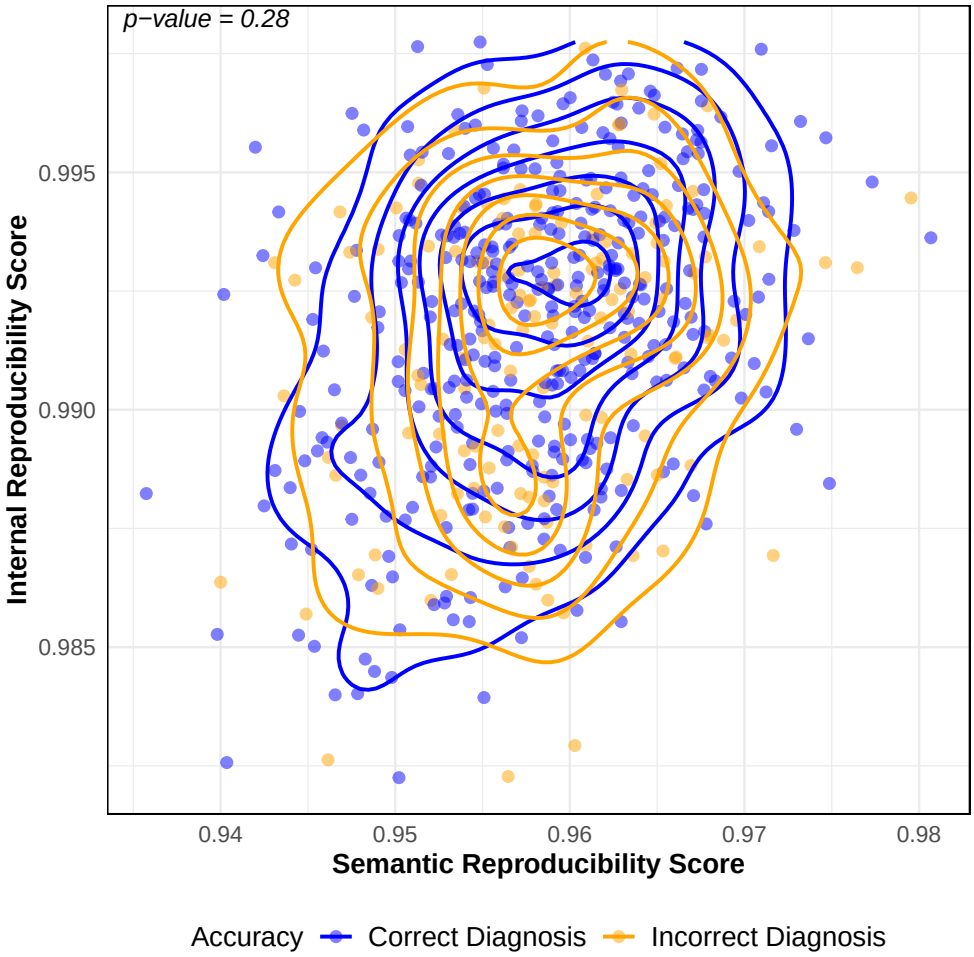

Fig. 5: **Semantic and Internal Reproducibility Scores for ChatGPT-4 on the MedQA (U.S. Medical Licensing Examination) dataset stratified by prompt.** Each dot represents a single clinical case evaluated by ChatGPT-4, with Semantic Reproducibility Score on the x-axis and Internal Reproducibility Score on the y-axis. Contour lines represent two-dimensional density estimates for cases that were correctly diagnosed by ChatGPT-4 (blue) and incorrectly diagnosed by ChatGPT-4 (orange). P-values are calculated based on a two-sided multivariate Kruskal-Wallis test at the 0.05 level. Higher values indicate greater reproducibility for both metrics.

4 DISCUSSION

We developed a regulatory-informed statistical framework for quantifying the repeatability and reproducibility of LLM outputs. Because the framework is agnostic to the specific LLM, prompting strategy, and evaluation task, it can be applied across diverse biomedical use cases. In practice, researchers can use the repeatability and reproducibility metrics to systematically compare output variability across different LLMs, prompts, model configurations, or evaluation datasets. Rather than defining universal thresholds, these metrics are primarily intended to support relative comparisons. While application-specific thresholds for these metrics may be defined in specific use cases, establishing such thresholds is beyond the scope of this study.

From our empirical evaluation, three key findings emerged. First, repeatability and reproducibility scores varied less for UDN cases than USMLE questions. One possible explanation is that the longer and more detailed narrative structure of real-world patient cases may constrain the range of plausible model responses, resulting in less variability across prompts. Additional studies across other clinical datasets are necessary to determine whether this finding generalizes beyond the specific datasets examined here. Second, prompts invoking Bayesian diagnostic reasoning yielded higher Semantic Repeatability Scores for ChatGPT-4. This finding suggests that repeatability may depend not only on the model itself, but also on the prompting strategy used to elicit reasoning. These findings indicate that repeatability and reproducibility are not fixed properties of an LLM, but depend on the interplay among the model, prompting strategy, and evaluation dataset. Therefore, evaluations based on a single prompting strategy or benchmark dataset may not fully characterize model performance across different settings. Third, repeatability and reproducibility were generally not associated with diagnostic accuracy in our empirical analysis. This observation highlights that correctness and consistency are distinct aspects of model performance. A model may produce a correct output in a single run but fail to reproduce that answer consistently across repeated runs, or conversely produce consistent outputs that are incorrect. Therefore, evaluating repeatability and reproducibility alongside accuracy may provide a more comprehensive characterization of LLM performance.

This work is complementary to prior efforts in natural language processing that quantify and improve LLMs' output consistency. Raj *et al.* proposed semantic consistency metrics and demonstrated that consistency and accuracy are independent properties, an observation that aligns with our findings [39]. Similarly, Cui *et al.* introduced a divide-and-conquer approach to improve sentence-level consistency, demonstrating that consistency metrics can serve not only as evaluation tools but also as mechanisms for improving model behvaior [40]. Wang *et al.* proposed MONITOR, a framework for assessing factual reliability under prompt variability, further underscoring the limitations of accuracy as the standalone evaluation metric [41]. Complementing these general-domain studies, our work operationalizes repeatability and reproducibility using definitions from FDA recommendations on AI-enabled devices [16], providing a step toward regulatory-informed evaluation of LLMs.

This work has several limitations. First, due to the substantial computational cost of generating repeated outputs across prompts, models, and clinical cases, our empirical evaluation did not exhaustively cover all possible configurations. While we included

diverse models, prompts, and datasets to approximate real-world clinical variation, additional contexts should be explored in future work. Second, the Internal Repeatability and Reproducibility Scores rely on access to token-level probabilities. While these probabilities are available in autoregressive LLMs (e.g., ChatGPT, LLaMA), they may not be directly accessible for other LLM architectures. Therefore, the internal metrics proposed in this study are primarily applicable to autoregressive LLMs, whereas the semantic metrics are broadly applicable across LLM architectures. Future work includes how variability metrics can be integrated with human-centered evaluation approaches, such as clinician review or usability assessments, to better understand how variability in model outputs influences clinical interpretation and decision-making [42].

Beyond evaluating model performance, quantifying variability is also important when LLM outputs are treated as study endpoints in biomedical research. Incorporating repeatability and reproducibility into evaluation frameworks may support more rigorous experimental design, model comparison, and interpretation of LLM-generated outputs. Practical considerations, including computational cost, scalability, and integration into clinical workflows, will influence how such evaluation frameworks are applied in practice [4, 42, 43].

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.* [33]'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 and K99LM014428 from the National Library of Medicine, and the Potocsnak Center for Undiagnosed and Rare Disorders.

Author Contributions

Conceptualization: C.S.
Data curation: C.S., B.R., C.H., R.J.T., T.A.C., R.H., L.B.
Formal analysis and interpretation: C.S., B.R., C.H., C.Y., A.W., L.B., J.F.P., B.A.M., H.X.
Methodology: C.S., B.R., C.H.
Writing – original draft: C.S.
Writing – review and editing: C.S., B.R., C.H., C.Y., T.J.T., T.A.C., R.H., A.W., L.B., J.F.P., B.A.M., H.X.
Supervision: H.X.

Project administration: C.S., H.X.
Funding acquisition: C.S., R.H.

References

[1] 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)

[2] 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)

[3] Woo, B.F.Y., Cato, K., Cho, H., You, S.B., Song, J.: The use of large language models in clinical documentation: A scoping review. International Journal of Nursing Studies, 105322 (2025)

[4] Hu, Y., Zuo, X., Zhou, Y., Peng, X., Huang, J., Keloth, V.K., Zhang, V.J., Weng, R.-L., Shyr, C., Chen, Q., et al.: Information extraction from clinical notes: are we ready to switch to large language models? Journal of the American Medical Informatics Association, 213 (2026)

[5] Ong, J.C.L., Jin, L., Elangovan, K., San Lim, G.Y., Lim, D.Y.Z., Sng, G.G.R., Ke, Y.H., Tung, J.Y.M., Zhong, R.J., Koh, C.M.Y., et al.: Large language model as clinical decision support system augments medication safety in 16 clinical specialties. Cell Reports Medicine **6**(10) (2025)

[6] Huang, J., Yang, D.M., Rong, R., Nezafati, K., Treager, C., Chi, Z., Wang, S., Cheng, X., Guo, Y., Klesse, L.J., *et al.*: A critical assessment of using chatgpt for extracting structured data from clinical notes. NPJ digital medicine **7**(1), 106 (2024)

[7] Busch, F., Hoffmann, L., Rueger, C., Van Dijk, E.H., Kader, R., Ortiz-Prado, E., Makowski, M.R., Saba, L., Hadamitzky, M., Kather, J.N., *et al.*: Current applications and challenges in large language models for patient care: a systematic review. Communications Medicine **5**(1), 26 (2025)

[8] Shyr, C., Cassini, T.A., Tinker, R.J., Byram, K.W., Embí, P.J., Bastarache, L., Peterson, J.F., Xu, H., Hamid, R., *et al.*: Large language models for rare disease diagnosis at the undiagnosed diseases network. JAMA Network Open **8**(8), 2528538–2528538 (2025)

[9] Shyr, C., Hu, Y., Tinker, R.J., Cassini, T.A., Byram, K.W., Hamid, R., Fabbri, D.V., Wright, A., Peterson, J.F., Bastarache, L., et al.: An artificial intelligence framework for end-to-end rare disease phenotyping from clinical notes using large language models. arXiv preprint arXiv:2602.20324 (2026)

[10] Gaber, F., Shaik, M., Allega, F., Bilecz, A.J., Busch, F., Goon, K., Franke, V.,

Akalin, A.: Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. npj Digital Medicine **8**(1), 263 (2025)

[11] Levra, A.G., Gatti, M., Mene, R., Shiffer, D., Costantino, G., Solbiati, M., Furlan, R., Dipaola, F.: A large language model-based clinical decision support system for syncope recognition in the emergency department: A framework for clinical workflow integration. European journal of internal medicine **131**, 113–120 (2025)

[12] Lu, Z., Peng, Y., Cohen, T., Ghassemi, M., Weng, C., Tian, S.: Large language models in biomedicine and health: current research landscape and future directions. Journal of the American Medical Informatics Association **31**(9), 1801–1811 (2024)

[13] Du, X., Zhou, Z., Wang, Y., Chuang, Y.-W., Li, Y., Yang, R., Zhang, W., Wang, X., Chen, X., Guan, H., et al.: Testing and evaluation of generative large language models in electronic health record applications: a systematic review. Journal of the American Medical Informatics Association, 233 (2026)

[14] Holtzman, A., Buys, J., Du, L., Forbes, M., Choi, Y.: The curious case of neural text degeneration. arXiv preprint arXiv:1904.09751 (2019)

[15] Tversky, A., Kahneman, D.: The framing of decisions and the psychology of choice. science **211**(4481), 453–458 (1981)

[16] 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

[17] 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)

[18] Vasey, B., Nagendran, M., Campbell, B., Clifton, D.A., Collins, G.S., Denaxas, S., Denniston, A.K., Faes, L., Geerts, B., Ibrahim, M., et al.: Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: Decide-ai. bmj **377** (2022)

[19] Liu, X., Rivera, S.C., Moher, D., Calvert, M.J., Denniston, A.K., Ashrafian, H., Beam, A.L., Chan, A.-W., Collins, G.S., Deeks, A.D.J., *et al.*: Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the consort-ai extension. The Lancet Digital Health **2**(10), 537–548 (2020)

[20] Kalf, A.J., Spruijt-Metz, D.: Variation in diagnoses: influence of specialists' training on selecting and ranking relevant information in geriatric case vignettes. Social Science & Medicine **42**(5), 705–712 (1996)

[21] Benner, P., Hughes, R.G., Sutphen, M.: Clinical reasoning, decisionmaking, and action: Thinking critically and clinically. Patient safety and quality: An evidence-based handbook for nurses (2008)

[22] Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)

[23] Lin, C.-Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81 (2004)

[24] Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002)

[25] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., *et al.*: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)

[26] Wang, Z., Wu, D., Nguyen, Q., Wang, K.: Integrating chain-of-thought and retrieval augmented generation enhances rare disease diagnosis from clinical notes. Medicine Bulletin (2025)

[27] Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., Sontag, D.: Large language models are few-shot clinical information extractors. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 1998–2022 (2022)

[28] Liu, A.T., Xiao, W., Zhu, H., Zhang, D., Li, S.-W., Arnold, A.: Qaner: Prompting question answering models for few-shot named entity recognition. arXiv preprint arXiv:2203.01543 (2022)

[29] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171 (2022)

[30] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)

[31] Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)

[32] Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk,

J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)

[33] 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)

[34] 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)

[35] Gahl, W.A., Tifft, C.J.: The nih undiagnosed diseases program: lessons learned. Jama **305**(18), 1904–1905 (2011)

[36] Li, A., Qian, L., Du, M., Yin, Y., Hu, Y., Sun, Z., Fu, Y., Kim, H., Stutz, E., Ai, X., et al.: Memorization in large language models in medicine: Prevalence, characteristics, and implications. arXiv preprint arXiv:2509.08604 (2025)

[37] Splinter, K., Adams, D.R., Bacino, C.A., Bellen, H.J., Bernstein, J.A., Cheatle-Jarvela, A.M., Eng, C.M., Esteves, C., Gahl, W.A., Hamid, R., *et al.*: Effect of genetic diagnosis on patients with previously undiagnosed disease. New england journal of medicine **379**(22), 2131–2139 (2018)

[38] Macnamara, E.F., D'Souza, P., Network, U.D., Tifft, C.J.: The undiagnosed diseases program: Approach to diagnosis. Translational Science of Rare Diseases **4**(3-4), 179–188 (2020)

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

[40] 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 29th Conference on Empirical Methods in Natural Language Processing, pp. 334–361 (2024)

[41] 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 28th Conference on Empirical Methods in Natural Language Processing, pp. 15361–15371 (2023)

[42] Bakken, S.: Improved evaluation frameworks are required to move application of LLMs from research into clinical practice. Oxford University Press (2026)

[43] You, J., Shin, H.: Structural insights into clinical large language models and their barriers to translational readiness. Journal of the American Medical Informatics Association, 230 (2026)