arXiv:2506.15339v3 [cs.CL] 26 Feb 2026

DeVisE: Towards the Behavioral Testing of Medical Large Language Models

Camila Zurdo Tagliabue[1], **Heloisa Oss Boll**[1,2], **Aykut Erdem**[3], **Erkut Erdem**[4], **Iacer Calixto**[1,2]

[1]Department of Medical Informatics, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands
[2]Amsterdam Public Health, Methodology, Amsterdam, The Netherlands
[3]Koç University, Istanbul, Turkey [4]Hacettepe University, Ankara, Turkey

Abstract

Large language models (LLMs) are increasingly applied in clinical decision support, yet current evaluations rarely reveal whether their outputs reflect genuine medical reasoning or superficial correlations. We introduce DeVisE (**De**mographics and **Vi**tal **s**igns **E**valuation), a behavioral testing framework that probes fine-grained clinical understanding through controlled counterfactuals. Using intensive care unit (ICU) discharge notes from MIMIC-IV, we construct both raw (real-world) and template-based (synthetic) variants with single-variable perturbations in demographic (age, gender, ethnicity) and vital sign attributes. We evaluate eight LLMs, spanning general-purpose and medical variants, under zero-shot setting. Model behavior is analyzed through (1) input-level sensitivity, capturing how counterfactuals alter perplexity, and (2) downstream reasoning, measuring their effect on predicted ICU length-of-stay and mortality. Overall, our results show that standard task metrics obscure clinically relevant differences in model behavior, with models differing substantially in how consistently and proportionally they adjust predictions to counterfactual perturbations.[1]

1 Introduction

Large language models (LLMs) are increasingly applied in the medical domain and show strong performance across clinical tasks (Gu et al., 2021; McDuff et al., 2023; Van Veen et al., 2024; Singhal et al., 2025). However, conventional medical benchmarks (Xu et al., 2023; Bakhshandeh, 2023; Yao et al., 2024) offer limited insight into how clinically grounded is an LLM's manipulation of key clinical variables when making predictions, such as predicting the risk of mortality of a patient in the hospital (Van Aken et al., 2021a; MacPhail et al., 2024; Jullien et al., 2024; Van Veen et al., 2024; Singhal et al., 2025). A key question is whether an LLM makes use of clinical concepts similarly to how a human clinician would, or whether it relies on biases, shortcuts, and spurious correlations.

To illustrate how we address this issue, let us use the example of an *'original patient'* in which a clinically meaningful variable such as the `heart rate` is changed from 89 bpm ('normal') to 120 bpm ('high'), yielding a *'counterfactual patient'*. We build on the intuition that, in such cases, an LLM should both adjust 1) *the probability of the patient clinical profile*, such that the counterfactual patient profile should become *less likely the further it gets from the original*, and 2) *the probability of associated outcomes*, such as predicting a *higher risk of mortality* for the counterfactual compared to the original patient, since the clinical condition of the patient deteriorated when moving from `heart rate` 89 bpm ('normal') to 120 bpm ('high').

In this work, we propose **DeVisE**, a *clinically-grounded evaluation framework for medical LLMs based on behavioural testing and counterfactual evaluation*. We show an overview of DeVisE in Figure 1. (1) **Admission notes:** We create admission notes for intensive care unit (ICU) patients from MIMIC-IV (Johnson et al., 2023b) discharge summaries following Van Aken et al. (2021b); Röhr et al. (2024). We extract key clinical variables: demographics (`age`, `gender`, `ethnicity`) from MIMIC's structured data, and vital signs (`heart rate`, `systolic blood pressure`, `diastolic blood pressure`, `body temperature`, `oxygen saturation`, `respiratory rate`) from the admission notes. We choose these variables due to their relevance to the two downstream tasks we use in our evaluation (Hempel et al., 2023; Candel et al., 2022) as well as for their low missingness rate. We use both raw admission notes and a template-based version of the same notes whereby only the variables of interest are included. (2) **Counterfactual generation:** For each patient, we

[1]Our benchmark is available at https://github.com/camztag/DeVisE

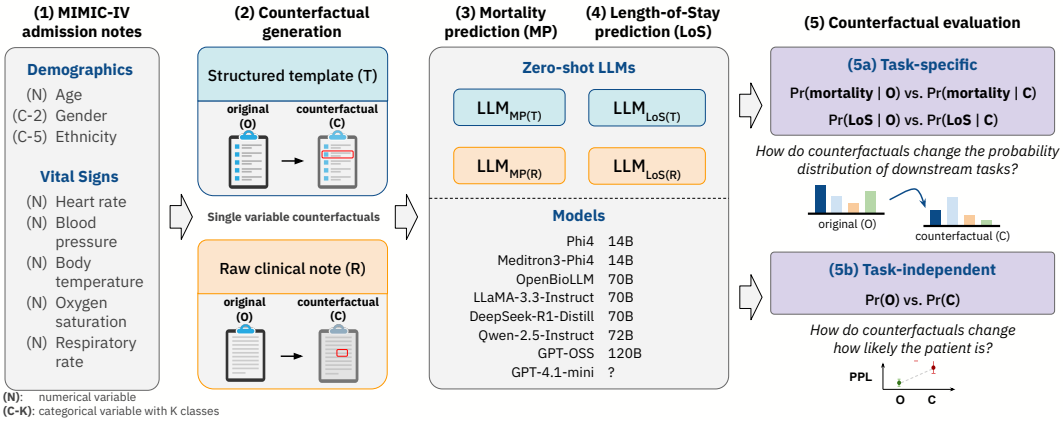


Figure 1: Overview of **DeVisE** in 5 steps: (1) We create admission notes using MIMIC-IV discharge summaries; (2) We create counterfactuals for template-based and raw notes; (3–4) We use original and counterfactual notes to predict mortality and LoS in a zero-shot setting, and (5) We evaluate and compare LLMs in (5a) a task-specific setting—on how counterfactuals affect mortality and LoS probabilities—and in (5b) a task-independent setting, by comparing how LLM perplexity change between original and counterfactual examples.

create single-variable counterfactuals for all variables, e.g., change a patient's `heart rate` from 89 to 120, while keeping everything else unchanged. (3–4) **Mortality** and **Length-of-stay prediction:** We apply LLMs in a zero-shot setting to predict two downstream tasks: mortality and length-of-stay (LoS) in the ICU. (5) **Counterfactual evaluation:** In (5a) **Task-specific evaluation**, we investigate how counterfactuals change the probability of the two downstream tasks. For counterfactuals implying a change in the patient condition, we expect the LLM to reflect that in the downstream probabilities. For instance, in our example the patient condition deteriorates with the `heart rate` changing from 89 / 'normal' to 120 / 'high'. We thus expect the LLM to predict a higher risk of mortality and longer LoS in the ICU for the counterfactual relative to the original patient. In (5b) **Task-independent evaluation**, we quantify to what extent counterfactuals affect how likely patients are, e.g., we compare the perplexity of the original vs. counterfactual patient notes. We expect that the further a counterfactual variable is from the true value of the clinical variable, the lower the probability the LLM should assign to the counterfactual.

Our main contributions are:

- We introduce **DeVisE** (**De**mographics and **Vi**tal **s**igns **E**valuation), a benchmark for clinical NLP that applies behavioral testing with *minimally differing counterfactuals* across demographic and vital sign variables.
- We release a dataset of $1,001$ raw and $1,001$ template-based validated notes, each with corresponding counterfactuals, enabling controlled evaluation of model robustness to clinically meaningful counterfactuals.
- We compare eight state-of-the-art LLMs across dimensions: template-based vs. raw clinical note, task-independent vs. task-specific performance, medical vs. general-purpose vs. reasoning-oriented LLMs.

While not exhaustive, DeVisE provides a controlled and clinically meaningful testbed for probing reasoning patterns, paving the way for broader evaluations and safer deployment in real-world medical settings.

2 Method

DeVisE is designed as a test-only clinical evaluation benchmark and includes three main tests, illustrated in Figure 1: two task-specific tests measuring how *mortality* and *LoS prediction* changes under counterfactuals, and one task-independent test measuring how *patient clinical profile probability* changes under counterfactuals. LLMs are not trained on original or counterfactual examples we generate in any way, and in that sense *DeVisE is a zero-shot benchmark*.

Background Let $\mathcal{P} \in \mathcal{D}$ denote the set of patients in a dataset $\mathcal{D}$ (detailed further in §2.1) and $\mathcal{V}$ the set of clinical variables (explained in §2.2). Each variable $v \in \mathcal{V}$ takes values in a domain $\mathcal{X}_v$ defined by ranges (numerical variables) or classes (categorical variables) defined in Tables 1–2. In §2.3 we explain how we map numerical variables $v \in \mathcal{V}$ onto *clinically-grounded severity scales* for

our analysis. Patient $i \in \mathcal{P}$ is represented by a feature vector $\mathbf{x}_i = (x_{iv})_{v \in \mathcal{V}} \in \mathcal{X} = \prod_{v \in \mathcal{V}} \mathcal{X}_v$.

For each patient i and variable v, let $\mathcal{A}_{iv} \subset \mathcal{X}_v$ denote the set of *alternative values* or *possible counterfactuals*, with $a \neq x_{iv}$ for all $a \in \mathcal{A}_{iv}$. A *single-variable counterfactual* for patient i is the feature vector obtained by replacing the original value of v in $\mathbf{x}_i$ with a, while keeping all other variables unchanged; we denote this vector by $\tilde{\mathbf{x}}_{\mathrm{i}}^{(\mathrm{v} \leftarrow \mathrm{a})}$. The set of all counterfactuals for patient i and variable v is $\mathcal{C}_{iv} = \{\tilde{\mathbf{x}}_{\mathrm{i}}^{(\mathrm{v} \leftarrow \mathrm{a})} \mid v \in \mathcal{V}, \forall a \in \mathcal{A}_{iv}\}$, and the set of all counterfactuals for all variables for a patient i is $\mathcal{C}_i = \{\tilde{\mathbf{x}}_{\mathrm{i}}^{(\mathrm{v} \leftarrow \mathrm{a})} \mid \forall v \in \mathcal{V}, \forall a \in \mathcal{A}_{iv}\}$. We explain how we generate counterfactuals in § 2.4 and detail our manual validation of the original and counterfactual patients' data in § 2.5.

Each patient i is associated with two categorical outcomes: mortality $y_i^{\text{mort}} \in \mathcal{Y}^{\text{mort}}$ and length-of-stay $y_i^{\text{los}} \in \mathcal{Y}^{\text{los}}$, discussed further in § 3.1.1 and § 3.1.2, respectively. Let $f_\theta : \mathcal{X} \to \{\mathcal{Y}^{\text{mort}}, \mathcal{Y}^{\text{los}}\}$ denote a model with parameters θ; we write $f_\theta(\mathbf{x}_i)$ and $f_\theta(\tilde{\mathbf{x}}_{\mathrm{i}}^{(\mathrm{v} \leftarrow \mathrm{a})})$ for predictions on observed and counterfactual instances, respectively.

2.1 Source dataset

We source our data from the MIMIC-IV database (Johnson et al., 2023b,a), which includes 65,366 intensive care unit (ICU) patients from the Beth Israel Deaconess Medical Center in Boston, MA, USA, and documents the full course of a hospital stay. MIMIC-IV includes patients' structured data (such as demographics, prescribed medication, vital signs, and lab results) and also free-text discharge notes. Discharge notes are semi-structured and include information about the patient medical condition prior to admission to the ICU, as well as about the patient's ICU stay. Our benchmark includes 1,001 randomly selected admission notes from MIMIC-IV's test set, representative of the adult population in MIMIC-IV: 45% female, mean age 64±17 years, and 31% non-white (see Table 4).

From discharge notes to admission notes Since in our task-specific tests we predict mortality and LoS in the ICU, we preprocess discharge summaries to remove all information pertinent to the current ICU stay. We follow Van Aken et al. (2021b); Röhr et al. (2024) and keep only information available at the time of admission in what we refer to as **admission notes**. Specifically, we preprocess and clean discharge summaries retaining only the following fields (and redacting everything else): *chief complaint*, *present illness*, *medical history*, *admission medications*, *allergies*, *physical exam*, *family history*, and *social history*.

Variable (**Range** $\mathcal{X}_v$)	**Range per class**	**Classes**
Age (18–91)	18–35	young adults
	36–55	middle aged adults
	56–75	older adults
	76–91	elderly
Gender (N/A)	—	female
	—	male
Ethnicity (N/A)	—	Asian & Pacific
	—	Black
	—	Hispanic/Latino
	—	Other/Unknown
	—	White

Table 1: Demographic variables used in DeVisE.

Raw admission notes vs. template-based notes Raw admission notes may include a wide range of information not directly relevant for particular downstream tasks. For that reason, we create a template-based version of a patient's admission note whereby only the variables we are interested in and their value are included (see § A.1 and § A.2 for examples of a raw note and its corresponding template-based version). Template-based notes allow us to isolate model responses to variable changes from confounding arising from natural variation and noise in the original raw text notes.

When using template-based notes, we shuffle the variables' serialisation order before feeding a note to an LLM. We discuss experiments with raw- and template-based notes in detail later in § 3.

2.2 Clinical variables

Demographics We use `age`, `gender`, and `ethnicity` as demographic variables (Table 1). We choose these variables for their relevance to studying model biases, their predictive association to our downstream tasks, and their availability (Van Aken et al., 2021a; Celi et al., 2022; Zhao et al., 2024). In MIMIC-IV demographic variables are redacted from discharge notes to de-identify protected health information; we thus extract these variables from the patient's structured data. There are no missing demographic variable.

Vital signs In Table 2 we show the vital sign variables we use in DeVisE: `heart rate` (`HR`), `systolic blood pressure` (`SBP`), `diastolic blood pressure` (`DBP`), `respiratory rate` (`RR`), `oxygen saturation` (`O2Sat`), and `body`

temperature (T). These are all clinically relevant indicators in early detection and prognosis across clinical specialties (Downey et al., 2017; Alghatani et al., 2021; Herasevich et al., 2022), and abundant in MIMIC-IV, having missing values ranging between 4% and 27%. We extract all vital sign variables directly from patients' admission notes using LLaMA-3.3-70B-Instruct with few-shot prompting (Grattafiori et al., 2024). For more details, please refer to § A.3.

2.3 Clinically-informed data preparation

We map numerical variables onto bins and explain how we do this for demographics and vital signs.

Demographic variables The only numerical demographic variable is age; gender and ethnicity are already categorical. We preprocess age by binning ages into four classes, as illustrated in Table 1.

Vital sign variables All vital signs are numerical variables and are mapped onto *clinically-grounded severity scales*. For a given variable $v \in \mathcal{V}$, these scales provide immediate clinical interpretation: normal values indicate an ideal state, high or low values reflect a deterioration in the patient's condition, and very high or very low values correspond to the most critical states. These severity scales are derived from established clinical guidelines for each vital sign and discretize continuous measurement into small set of interpretable ranges to enable consistent counterfactual analysis (Infectious Diseases Society of America, 2003; Royal College of Physicians, 2017; Society of Critical Care Medicine, 2015; World Health Organization, 2016; Society of Critical Care Medicine, 2021; American Heart Association, 2024). Oxygen saturation is a special case, as values above normal are not defined. For more details refer to § A.4.

2.4 Counterfactual generation

We generate counterfactual notes by modifying one clinical variable $v \in \mathcal{V}$ at a time—either a demographic attribute or a vital sign—while keeping the rest of the note unchanged. For each patient $i \in \mathcal{P}$, we uniformly draw five samples per class per variable $v \in \mathcal{V}$. Classes for demographic variables are defined in Table 1 and classes for vital sign variables are the five severity bins in Table 2. We draw five counterfactuals uniformly from each class to ensure a balanced coverage across all classes. In total, we generate **166,731 counterfactuals**: 20,020 (age), 1,001 (gender), 4,004 (ethnicity), 30,895 (SBP), 30,895 (DBP), 27,070 (HR), 20,590 (T), 19,404 (RR), and 12,852 (O2Sat).

2.5 Automatic and human validation

We ensure DeVisE's data quality (including counterfactuals) through automatic and manual checks.

Variable extraction and replacement As mentioned earlier, demographic variables are extracted from patient structured data, whereas vital sign variables are extracted from admission notes using LLaMA-3.3-70B-Instruct with few-shot prompting. For each variable replacement, we first use the GNU diff tool[2] to verify that edits affect only the intended variable, flagging anomalies for human review. We manually validated all the raw and template-based notes from original and counterfactual data in DeVisE and observed $\sim 5\%$ error rate in vitals extracted with LLMs, while demographic replacements drawn from structured data showed no error. All errors were manually corrected.

3 Experimental Setup

We now describe how we assess LLMs' sensitivity and robustness to counterfactuals. Towards this purpose, we describe our evaluation protocol and experiments using two downstream tasks, mortality and length-of-stay prediction (§ 3.1), and also independently of any downstream task (§ 3.2). We provide details on the multiple LLMs in § 3.3.

3.1 Task-specific evaluation

Inclusion and exclusion criteria We perform mortality and length-of-stay prediction in the ICU within the first 24 hours. We only use the first ICU stay for the patients in our cohort and, for patients with multiple ICU stays: when these occur less than 48 hours apart, we merge them into a single ICU stay; when ICU stays are separated by more than 48 hours, we treat them as distinct ICU admissions and keep only the first admission. We do not include patients who die within the first 24 hours of admission to the ICU or who have an ICU stay shorter than 24 hours.

3.1.1 Mortality classification

We frame mortality prediction as binary classification with labels $\mathcal{Y}^{\text{mort}} \in \{0, 1\}$ indicating if the patient dies (1) or not (0) during the hospital admission. The mortality rate in our cohort is 7%.

[2] https://www.gnu.org/software/diffutils/

Vital Sign (Range $\mathcal{X}_v$)	Very low (-2)	Low (-1)	Normal (0)	High (+1)	Very high (+2)	Missing (%)
Heart Rate (1–200 bpm)	1–40	41–50	51–90	91–110	111–200	5.7%
Systolic Blood Pressure (1–220 mmHg)	1–70	71–89	90–119	120–139	140–220	4.3%
Diastolic Blood Pressure (1–140 mmHg)	1–40	41–59	60–79	80–89	90–140	4.3%
Respiration Rate (1–50 bpm)	1–8	9–11	12–20	21–24	25–50	12%
Oxygen Saturation (1–100 %SpO_2)	1–91	92–95	96–100	—	—	5.1%
Temperature (70.0–110.0 ^{o}F)	70.0–89.4	89.5–94.9	95.0–100.2	100.3–103.9	104.0–110.0	27%

Table 2: Clinically meaningful ranges and severity scores. Severity ranges are derived from clinical guidelines.

3.1.2 Length-of-Stay (LoS) classification

LoS measures how long a patient stays in the ICU and is defined as the number of hours between admission and discharge. LoS prediction is modelled as a 4-way classification task with labels $\mathcal{Y}^{\text{los}} \in \{0, 1, 2, 3\}$ based on training data quantiles: label 0 for $\texttt{LoS} < Q_{25}$ (24 to 39 hours), label 1 for $Q_{25} \leq \texttt{LoS} < Q_{50}$ (39 to 59 hours), label 2 for $Q_{50} \leq \texttt{LoS} < Q_{75}$ (59 to 112 hours), and label 3 for $\texttt{LoS} \geq Q_{75}$ (over 112 hours). LoS bins are evenly represented with each class accounting for ∼25% of the data.

3.1.3 Evaluation metrics

We use the following evaluation metrics for mortality (§ 3.1.1) and length-of-stay (§ 3.1.2) prediction.

Label probability shift For all variables $v \in \mathcal{V}$, KL is the KL divergence that measures how distant are label probability distributions for counterfactuals $\tilde{\mathbf{x}}_{\text{i}}^{(\text{v}\leftarrow\text{a})}$ relative to original patients $\mathbf{x}_{iv}$.

$$\texttt{KL} = \mathop{\mathbb{E}}_{i\in\mathcal{P},\, \tilde{\mathbf{x}}_{\text{i}}^{(\text{v}\leftarrow\text{a})}\in\mathcal{C}_i} \Big[g(\tilde{\mathbf{x}}_{\text{i}}^{(\text{v}\leftarrow\text{a})}, \mathbf{x}_{iv})\Big], \quad (1)$$
$$g(a, b) = \text{KL}(f_\theta(Y|a) \parallel f_\theta(Y|b)),$$

where f_θ is an LLM with parameters θ that computes the probability distribution over labels Y.

Percentage of label flips For all variables $v \in \mathcal{V}$, we compute the percentage of counterfactuals which induce a flip in the highest probability label. This metric measures prediction stability in terms of class labels and indicates whether counterfactual perturbations caused categorical prediction shifts.

$$\texttt{Flips(\%)} = \mathop{\mathbb{E}}_{i\in\mathcal{P},\, \tilde{\mathbf{x}}_{\text{i}}^{(\text{v}\leftarrow\text{a})}\in\mathcal{C}_i} \Big[h(\tilde{\mathbf{x}}_{\text{i}}^{(\text{v}\leftarrow\text{a})}, \mathbf{x}_{iv})\Big], \quad (2)$$

$$h(a, b) = \begin{cases} 1 : \text{argmax} f_\theta(Y|a) \neq \\ \qquad\qquad \text{argmax} f_\theta(Y|b) \\ 0 : \text{otherwise} \end{cases}$$

where f_θ is an LLM with parameters θ that computes the probability distribution over labels Y.

We approximate the expectations in Eqs. (1–2) using counterfactual samples described in § 2.4.

Correct direction In addition to probability shift and label stability, we assess whether counterfactual perturbations induce prediction changes consistent with the expected direction of clinical severity. For a counterfactual $\tilde{\mathbf{x}}_{\text{i}}^{(\text{v}\leftarrow\text{a})} \in \mathcal{C}_{iv}$ for a patient i and variable v, we define the severity shift as

$$\Delta s_{iv} = |\text{sev}(\tilde{\mathbf{x}}_{\text{i}}^{(\text{v}\leftarrow\text{a})})| - |\text{sev}(\mathbf{x}_i)|, \quad (3)$$

where $\text{sev}(\cdot)$ denotes the severity scale associated with the variable value. A positive shift ($\Delta s_{iv} > 0$) corresponds to increased severity, while a negative shift ($\Delta s_{iv} < 0$) corresponds to decreased severity.

We denote the expected LoS for a patient i by

$$\mathbb{E}[y_i^{\text{los}} \mid \mathbf{x}_i] = \sum_{l\in\mathcal{Y}^{\text{los}}} f_\theta(l \mid \mathbf{x}_i) \cdot \mu_l,$$

where μ_l is the empirical mean LoS for bin l computed over the training data. Similarly, we denote the expected mortality for a patient i simply by

$$\mathbb{E}[y_i^{\text{mort}} \mid \mathbf{x}_i] = f_\theta(y_i^{\text{mort}}{=}1 \mid \mathbf{x}_i).$$

Finally, we consider downstream predictions for a counterfactual as directionally correct iif

$$\text{s}\Big(\mathbb{E}[y_i \mid \tilde{\mathbf{x}}_{\text{i}}^{(\text{v}\leftarrow\text{a})}] - \mathbb{E}[y_i \mid \mathbf{x}_i]\Big)\Big) = \text{s}(\Delta s_{iv}), \quad (4)$$

where $y_i \in \{y_i^{\text{los}}, y_i^{\text{mort}}\}$ denotes the expected LoS and mortality probability, respectively, and $\text{s}(\cdot)$ is the sign function. We report %CD as the proportion of counterfactuals that satisfy Equation (4).

Monotonicity For each severity scale $s \in \{-2, -1, 0, +1, +2\}$ (see Table 2), we compute the mean of the differences between expected downstream predictions for original and counterfactual patients as below.

$$m(s) = \mathbb{E}[y_i \mid \tilde{\mathbf{x}}_{\text{i}}^{(\text{v}\leftarrow\text{a})}] - \mathbb{E}[y_i \mid \mathbf{x}_i], \quad (5)$$

where $y_i \in \{y_i^{\text{los}}, y_i^{\text{mort}}\}$ again denotes the expected LoS and mortality probability, respectively. We compute $m(s)$ for each s by aggregating all original–counterfactual pairs $(\mathbf{x}_i, \tilde{\mathbf{x}}_i^{(\text{v}\leftarrow\text{a})})$ with a corresponding $\Delta s_{iv} = s$.

We consider predictions as monotonic iff all the inequalities below hold.

$$m(s-1) \le m(s) \le m(s+1), \quad \text{and} \quad (6)$$

$$\text{if}(s \neq 0) \begin{cases} s \in \{-2,-1\} : m(s) < 0 \\ s \in \{+1,+2\} : m(s) > 0 \end{cases}$$

We report %Mono as the proportion of predictions satisfying both conditions.

3.2 Task-independent evaluation

Here we analyze how the probability an LLM assigns to the original patient changes under a counterfactual *independently of any task*.

3.2.1 Evaluation metrics

We first note that we use perplexity as a proxy for the probability of an admission note as below.

$$\text{PPL} = \exp\left(-\tfrac{1}{N}\textstyle\sum_{n=1}^{N} \log P(t_n)\right),$$

where t_n are the tokens in a patient note.

Patient probability shift We calculate the expected shift in perplexity ΔPPL brought by a counterfactual relative to an original patient as below.

$$\Delta\text{PPL} = \mathop{\mathbb{E}}_{i\in\mathcal{P},\, \tilde{\mathbf{x}}_i^{(\text{v}\leftarrow\text{a})}\in\mathcal{C}_{iv}} \left[\text{PPL}_{\tilde{\mathbf{x}}_i^{(\text{v}\leftarrow\text{a})}} - \text{PPL}_{\mathbf{x}_{iv}}\right], \quad (7)$$

where $\text{PPL}_{\tilde{\mathbf{x}}_i^{(\text{v}\leftarrow\text{a})}}$ and $\text{PPL}_{\mathbf{x}_{iv}}$ are the perplexities of counterfactual and original patient admission notes, respectively.

We approximate the expectation in Eq. (7) using counterfactual samples described in § 2.4.

3.3 Large language models (LLMs)

We analyze the performance of 8 LLMs via in-context learning (zero-shot) for downstream tasks. For our experiments, we choose LLMs of varying sizes, architectures, domains, and reasoning capabilities: **Phi4-14B** (Abdin et al., 2024), **Meditron3-Phi4-14B** (OpenMeditron, 2025), **LLaMA-3.3-Instruct-70B** (Grattafiori et al., 2024), **OpenBioLLM-70B** (Ankit Pal, 2024), **DeepSeek-R1-Distill-70B** (Guo et al., 2025), **Qwen-2.5-Instruct-72B** (Team, 2024), **GPT-OSS-120B** (OpenAI, 2025), and **GPT-4.1-mini** (OpenAI(gpt-4.1-mini), 2025).[3] Zero-shot prompts we use are available in § A.6.

We analyze LLMs along three axes: (1) *template vs. raw admission notes* (2) *task-independent vs. task-specific* (3) *medical-domain vs. general-purpose vs. reasoning-oriented*. These comparisons reveal how model specialization and reasoning capacity affect robustness, sensitivity, and behavioral consistency. We summarise all LLM specifications in Table 5, § A.7.1.

[3]We test GPT 4.1 mini within a private and secure environment according to the MIMIC-IV data usage agreement.

4 Results

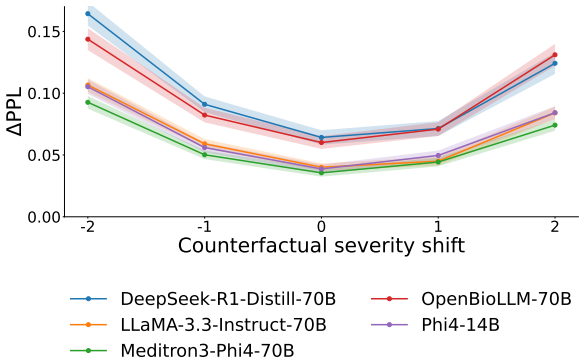
Below we report results for our task-independent (§ 4.1) and task-specific (§ 4.2) evaluations.

4.1 Task-independent

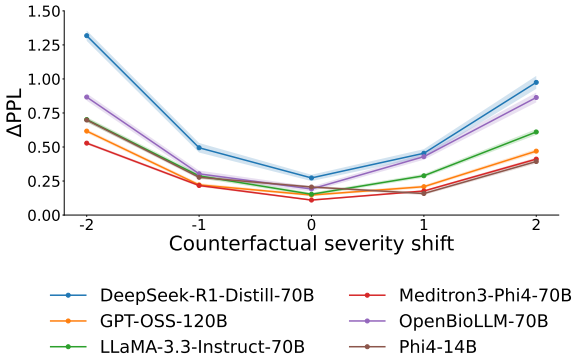
We observe small behavioral differences in terms of change in perplexity across models that scale scale with severity (Figure 2). These differences are consistent in both note modalities, being effect on template-based notes being more pronounced than in raw notes. This means that models are linguistically *surprised* by counterfactuals even when not trained for a downstream task. GPT-OSS-120B shows a substantially larger ΔPPL than the other models, suggesting heightened sensitivity or instability to counterfactual perturbations.

4.2 Task-specific

Vital signs. We observe behavioral differences in the presence of downstream tasks (Figure 3). As with perplexity, KL also often scales with severity shifts in several models. However, the behaviors are more heterogeneous, GPT-4.1-mini presenting a monotonic increase, and Meditron3-Phi4-14B and OpenBioLLM-70B almost flat response. Moreover, GPT-OSS-120B again exhibits substantially a larger KL shift than the other models. LLM performance on template-based data, in both tasks, is comparable to standard machine learning models (i.e., see Table 3 and Table 6 in § A.7.2). For mortality prediction, class imbalance impacts the performance of both classical machine learning and LLM-based models. Overall, we note that using templates 1) improves accuracy, 2) increases the percentage of correct directional responses across both mortality and LoS tasks (Table 3), and 3) amplifies behavioral sensitivity in terms of KL and categorical prediction shifts. Medical and reasoning models often show lower KL, suggesting a more conservative overall behavior. Moreover, general-purpose models achieve higher task performance



(a) Raw notes.



(b) Template-based notes.

Figure 2: Average per-token ΔPPL across counterfactual severity shifts. ΔPPL grows with both increasing and decreasing severity, indicating consistent linguistic sensitivity. In (a) GPT-OSS-120B was excluded due to its substantially higher ΔPPL (2.5 ± 5.1). We observe the same pattern in (a) and (b) with higher effects in template notes. Except GPT-OSS-120B that inversely, presents a smaller response in raw notes. GPT-4.1-mini is omitted because we could not obtain per-token log probabilities for this model.

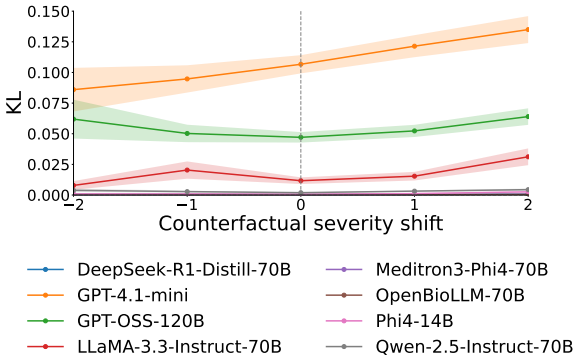


Figure 3: Average KL across counterfactual severity shifts for LoS task (raw notes). Models show different behaviours. GPT-4.1-mini, GPT-OSS-120B and LLaMA-3.3-Instruct-70B show larger KL shifts than other models.

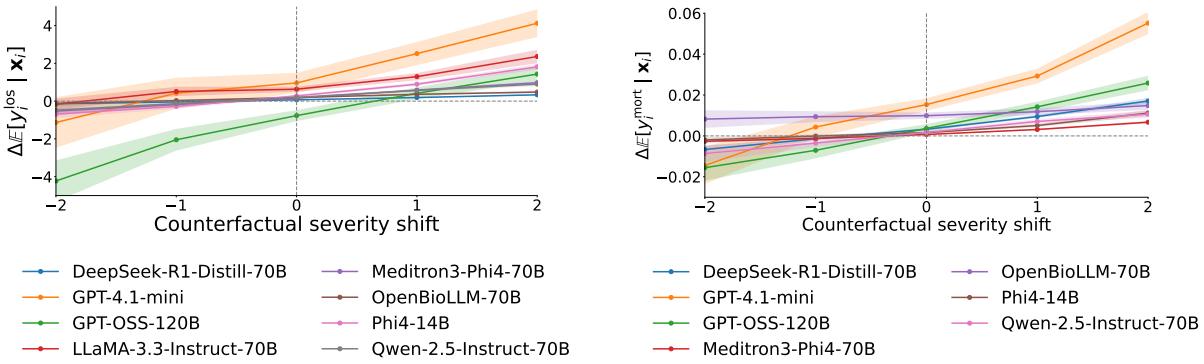
(especially on mortality) and exhibit greater sensitivity. GPT-OSS-120B shows consistently better performance and higher sensitivity to perturbations.

Demographics. Demographic perturbations also yield consistent, statistically significant effects (Figure 5). As expected, age shows the strongest influence: older age groups are associated with longer predicted LoS across models, even under minimal textual edits.

Gender and race effects are smaller on average but still affect boundary decisions, with flip rates `Flips(%)` up to 23%. Although most demographic comparisons are statistically significant, "Older adults" age for several models, and "White" for DeepSeek-R1-Distill-70B, and "Other/Unknown", "Black" and "Asian and Pacific" races for GPT-OSS-

Model	Acc.	F1	ΔKL±std	%CD	%Flip	%Mono
Mortality prediction, raw notes						
Phi4-14B	0.28	0.26	0.00 ± 0.01	**73.1**	0.9	**100.0**
Meditron3-Phi4-14B	0.41	0.36	0.00 ± 0.00	69.8	0.6	**100.0**
DeepSeek-R1-Distill-70B	0.60	0.48	0.00 ± 0.01	72.4	1.8	**100.0**
LLaMA-3.3-Instruct-70B	0.71	0.54	**0.97 ± 2.31**	69.5	**15.0**	62.5
OpenBioLLM-70B	0.82	0.62	0.01 ± 0.02	54.8	4.0	75.0
Qwen-2.5-Instruct-72B	0.67	0.52	0.01 ± 0.04	69.7	1.4	**100.0**
GPT-OSS-120B	**0.88**	**0.63**	0.06 ± 0.12	64.6	4.1	**100.0**
GPT-4.1-mini	0.62	0.50	0.18 ± 0.32	62.9	7.1	87.5
Average	0.62	0.49	0.15 ± 0.35	67.1	4.4	90.6
Mortality prediction, template-based notes						
Phi4-14B	0.73	0.53	1.37 ± 1.07	90.1	32.4	**100.0**
Meditron3-Phi4-14B	0.84	0.57	0.22 ± 0.14	**90.9**	26.6	**100.0**
DeepSeek-R1-Distill-70B	0.81	0.54	0.36 ± 0.15	90.7	25.0	**100.0**
LLaMA-3.3-Instruct-70B	0.78	0.53	3.41 ± 1.93	88.3	29.5	87.5
OpenBioLLM-70B	0.92	0.58	0.17 ± 0.08	82.4	10.4	87.5
Qwen-2.5-Instruct-72B	0.89	**0.59**	1.11 ± 0.50	89.7	16.3	**100.0**
GPT-OSS-120B	**0.93**	0.48	0.21 ± 0.24	82.4	1.6	75.0
GPT-4.1-mini	0.83	0.56	**3.86 ± 2.43**	88.3	**33.3**	87.5
Average	0.84	0.55	1.34 ± 0.82	87.9	21.9	92.2
Length-of-stay prediction, raw notes						
Phi4-14B	0.26	0.19	0.00 ± 0.01	**71.1**	1.6	**100.0**
Meditron3-Phi4-14B	0.27	0.21	0.00 ± 0.00	69.8	1.9	**100.0**
DeepSeek-R1-Distill-70B	0.28	0.17	0.00 ± 0.00	68.4	0.8	**100.0**
LLaMA-3.3-Instruct-70B	**0.31**	**0.28**	0.02 ± 0.07	70.5	2.3	87.5
OpenBioLLM-70B	0.29	0.25	0.00 ± 0.00	62.5	2.0	87.5
Qwen-2.5-Instruct-72B	0.30	0.21	0.00 ± 0.01	65.6	1.3	**100.0**
GPT-OSS-120B	0.28	0.19	0.06 ± 0.13	57.6	6.1	**100.0**
GPT-4.1-mini	0.30	0.25	**0.12 ± 0.23**	56.2	**11.3**	87.5
Average	0.29	0.22	0.03 ± 0.06	65.2	3.4	95.3
Length-of-stay prediction, template-based notes						
Phi4-14B	0.23	0.19	0.62 ± 0.43	89.5	36.9	**100.0**
Meditron3-Phi4-14B	0.23	0.20	0.13 ± 0.06	**91.2**	38.1	**100.0**
DeepSeek-R1-Distill-70B	0.27	0.16	0.03 ± 0.02	77.5	9.1	**100.0**
LLaMA-3.3-Instruct-70B	0.26	0.17	3.23 ± 1.53	88.3	32.9	87.5
OpenBioLLM-70B	0.25	0.15	0.02 ± 0.01	81.5	19.0	87.5
Qwen-2.5-Instruct-72B	**0.28**	**0.22**	0.87 ± 0.44	81.5	29.3	**100.0**
GPT-OSS-120B	**0.28**	0.17	0.51 ± 0.24	51.2	27.2	75.0
GPT-4.1-mini	0.24	0.17	**3.49 ± 1.59**	86.9	**39.8**	87.5
Average	0.26	0.18	1.11 ± 0.54	81.0	29.0	92.2

Table 3: Performance and behavioral metrics across note types and tasks (excluding precision/recall; see Table 7). Orange: medical LLM; blue: reasoning-oriented; white: general-purpose; gray: averages. Highest values are bolded and lowest underlined within each block.



(a) Length-of-stay (LoS).

(b) Mortality.

Figure 4: Expected $\Delta\mathbb{E}(y_i^{\text{los}}|\cdot)$ in hours (a) and probability of mortality $\Delta\mathbb{E}(y_i^{\text{mort}}|\cdot)$ (b) as a function of counterfactual severity shift. Positive severity shifts are expected to increase predicted LoS and mortality risk, while negative shifts are expected to decrease them. Most models follow a monotonic trend, indicating clinically aligned responses to vital sign counterfactuals. LLaMA-3.3-Instruct-70B was omitted due to its substantially larger response that dominates the scale, see § A.7.2.

120B, do not differ significantly from the remaining demographic groups.

Across models, DeepSeek-R1-Distill-70B and Qwen-2.5-Instruct-72B remain the most robust models, showing the lowest $\Delta\mathbb{E}(y_i^{\text{los}}|\cdot)$ and smaller flip rates. In contrast, Phi4-14B and GPT-OSS-120B produce the largest shifts for race and gender. Race-related effects are slightly larger on average than gender effects, and when significant, most models predict shorter LoS for female, Asian and Pacific, and White patients, and longer LoS for Black patients, aligning with known healthcare disparities (Macias-Konstantopoulos et al., 2023).

5 Related Work

LLMs are increasingly used in clinical settings, motivating a growing body of work on their robustness, fairness, and interpretability. Lee et al. (2025) analyzed model robustness to distribution shifts and missing data in patient triage, finding that while LLMs outperform traditional models, they exhibit demographic biases. Van Aken et al. (2021a) similarly showed that models with comparable AUROC scores can behave differently when evaluated for sensitivity to age, gender, and ethnicity. Zack et al. (2024) and Zhao et al. (2024) further demonstrated that GPT-4 and other LLMs often favor majority groups and yield less accurate predictions for minorities, reinforcing existing disparities.

Beyond fairness, several studies probe robustness through controlled input perturbations. MacPhail et al. (2024) introduced template-based tests for adverse drug event classification, revealing inconsistent performance even among models with similar discrimination. Kougia et al. (2024) showed that biomedical LLMs frequently fail at ordering events, compromising reliability in decision support. Likewise, Altinok (2024) and Aguiar et al. (2024) employed natural Language inference-based consistency and faithfulness, uncovering substantial variance across closely related models.

While these studies highlight the limits of traditional evaluation, most rely on structured templates, leaving raw clinical text largely unexamined. *Behavioral testing* offers a complementary approach, constructing minimally differing counterfactuals and checking the consistency of model input-output responses (Beizer and Wiley, 1996). Originally proposed in software engineering, it has since been adapted to NLP (Ribeiro et al., 2020), vision-and-language (Parcalabescu et al., 2022), and video-and-language tasks (Kesen et al., 2024).

6 Discussion

Task-specific vs. Task-independent. Even without task supervision, models respond systematically to increasing counterfactual severity. Perplexity shifts increase according to severity levels (Fig. 2), indicating that models internally register perturbations as increasingly improbable language events. This supports the use of task-agnostic behavioral probes as an early diagnostic signal of reasoning sensitivity.

Template-based vs. Raw notes. Templates amplify sensitivity to vital sign changes (Table 3), as shown in higher KLs—e.g., 0.15 (raw) →

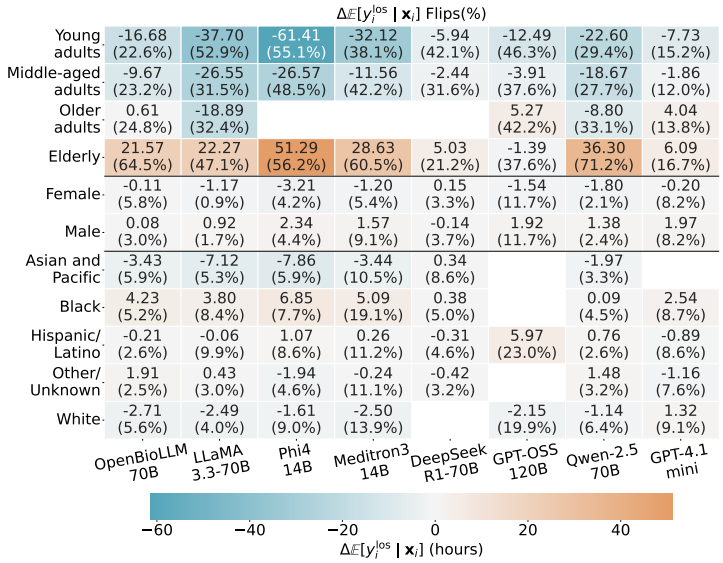


Figure 5: $\Delta\mathbb{E}(y_i^{\text{los}}|\cdot)$ and `Flips(%)` by demographics. Age produces the largest and most significant shifts, while race and gender effects are smaller but consistent. Blank cells denote non-significant t-test comparisons (p<0.05).

1.34 (templates)—and `Flips(%)`, e.g., 4.4→21.9% (mortality) and 3.4→29.0% (LoS). On average, templates increase correct directionality of predictions `%CD` (67.1→87.5% mortality, 66.5→79.9% LoS). We hypothesize that the absence of contextual information forces the model to rely more explicitly on vital signs. In contrast, raw notes with contextualized reasoning causes importance of vitals to be diluted by other information, potentially under-weighting important physiological changes.

Medical vs. General vs. Reasoning LLMs. Medical models show minimal sensitivity when evaluated on raw notes, likely reflecting conservative priors learned from biomedical corpora, and a preference for holistic interpretation of full clinical notes. Their consistently high `%Mono` indicates close alignment with clinical expectations. When applied to template-based notes their stability decrease and we observe higher `Flips(%)`, e.g., Meditron3-Phi4-14B achieves 38.1% for LoS.

General-purpose models show high sensitivity, which amplifies further under templates. However, this is not consistently accompanied by correct directional or monotonic reasoning, revealing instability despite competitive accuracy.

DeepSeek-R1-Distill-70B shows a modest sensitivity compared to medical models even in template settings, while maintaining high directionality and monotonicity. It favors smaller but calibrated shifts, consistent with expected physiological reasoning.

Demographic Variables. Demographic attributes subtly but consistently affect predictions. While mean LoS shifts for gender and race were small, elevated flip rates near decision boundaries (Fig. 5) raise fairness concerns. Consistent prediction patterns across gender and races for most models suggest that pretraining biases are pervasive, reinforcing the need for demographic auditing in clinical LLM evaluation.

7 Conclusion

In this work, we introduce DeVisE with clinically meaningful counterfactuals. We show that the behavioral testing of vital signs and demographic variables with counterfactuals provides insights into LLM behavior not captured by standard metrics. Vital sign perturbations consistently induce monotonic and directionally correct shifts in downstream predictions, indicating that models internalize physiological severity, but with substantial variation in sensitivity and stability across architectures. Demographic perturbations produce smaller yet systematic effects, even near decision boundaries, revealing persistent biases that align with known healthcare disparities.

Overall, our results highlight the importance of evaluating not only task performance, but how LLMs respond to clinically meaningful changes in patient information. We believe DeVisE is a step toward more robust, interpretable, and fairness-aware evaluation of clinical LLMs.

Limitations

Datasets and languages. We use MIMIC-IV, a large publicly available dataset including patients from a single hospital in the US. We see our work as an important step towards a clinically grounded behavioural testing of medical LLMs. We believe that important future work lies in further validating the generalisability of our framework in terms of other languages and healthcare systems.

Clinical variables. In this work, we do not cover an extensive range of model capabilities and focus instead on demographic and vital sign variables, focusing on demographic biases and assessing numerical reasoning. We leave other relevant aspects, such as evaluating the temporal reasoning of LLMs, to future work.

Missingness in clinical data. Clinical data is inherently affected by missingness, where absent measurements may reflect clinical practice. Although our analysis includes records with missing values and thus reflects realistic data conditions, explicitly assessing the impact of missingness versus observed inputs on model behavior was beyond the scope of this work.

Severity shifts and counterfactual sampling. Severity shifts are treated symmetrically in our analysis, such that very low and very high values correspond to the same absolute shift (+2), despite potentially differing clinical effects (e.g., on length of stay). Additionally, counterfactual values are sampled uniformly within severity bins, which can yield rare values in extreme bins. Future work should explore alternative sampling strategies better aligned with empirical distributions.

Number of counterfactuals. Some vital signs, such as oxygen saturation and respiration rate, had limited value ranges, resulting in fewer than five unique counterfactuals due to the small number of non-redundant available values.

Despite these limitations, our proposed framework is relevant and we see it as the first step towards clinically-grounded evaluation of LLMs for healthcare. While not exhaustive, our framework highlights biases and reasoning gaps paving the way to safe deployment of medical LLMs.

Acknowledgments

HOB and IC are funded by the project CaRe-NLP with file number NGF.1607.22.014 of the research programme AiNed Fellowship Grants which is (partly) financed by the Dutch Research Council (NWO). EE is funded by the project TUBITAK 2247-A National Outstanding Researchers Program Award No. 123C542.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

M. Aguiar, P. Zweigenbaum, and N. Naderi. 2024. Seme at semeval-2024 task 2: Comparing masked and generative language models on natural language inference for clinical trials. *arXiv preprint arXiv:2404.03977*.

Khalid Alghatani, Nariman Ammar, Abdelmounaam Rezgui, Arash Shaban-Nejad, and 1 others. 2021. Predicting intensive care unit length of stay and mortality using patient vital signs: machine learning model development and validation. *JMIR medical informatics*, 9(5):e21347.

D. Altinok. 2024. D-nlp at semeval-2024 task 2: Evaluating clinical inference capabilities of large language models. *arXiv preprint arXiv:2405.04170*.

American Heart Association. 2024. Understanding blood pressure readings and heart rate guidelines. https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings.

Malaikannan Sankarasubbu Ankit Pal. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B.

Sadra Bakhshandeh. 2023. Benchmarking medical large language models. *Nature Reviews Bioengineering*, 1(8):543–543.

Boris Beizer and J Wiley. 1996. Black box testing: Techniques for functional testing of software and systems. *iEEE Software*, 13(5):98.

Bart GJ Candel, Renée Duijzer, Menno I Gaakeer, Ewoud Ter Avest, Özcan Sir, Heleen Lameijer, Roger Hessels, Resi Reijnen, Erik W van Zwet, Evert de Jonge, and 1 others. 2022. The association between vital signs and clinical outcomes in emergency department patients of different age categories. *Emergency Medicine Journal*, 39(12):903–911.

Leo Anthony Celi, Jacqueline Cellini, Marie-Laure Charpignon, Edward Christopher Dee, Franck Dernoncourt, Rene Eber, William Greig Mitchell, Lama Moukheiber, Julian Schirmer, Julia Situ, and 1 others. 2022. Sources of bias in artificial intelligence that perpetuate healthcare disparities—a global review. *PLOS Digital Health*, 1(3):e0000022.

Candice L Downey, W Tahir, R Randell, JM Brown, and DG Jayne. 2017. Strengths and limitations of early warning scores: a systematic review and narrative synthesis. *International journal of nursing studies*, 76:106–119.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Lars Hempel, Sina Sadeghi, and Toralf Kirsten. 2023. Prediction of intensive care unit length of stay in the mimic-iv dataset. *Applied Sciences*, 13(12):6930.

Svetlana Herasevich, Kirill Lipatov, Yuliya Pinevich, Heidi Lindroth, Aysun Tekin, Vitaly Herasevich, Brian W Pickering, and Amelia K Barwise. 2022. The impact of health information technology for early detection of patient deterioration on mortality and length of stay in the hospital acute care setting: systematic review and meta-analysis. *Critical care medicine*, 50(8):1198–1209.

Infectious Diseases Society of America. 2003. Fever and neutropenia clinical practice guidelines. https://www.idsociety.org/practice-guideline/febrile-neutropenia/.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023a. MIMIC-IV-Note: Deidentified free-text clinical notes.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and 1 others. 2023b. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

M. Jullien, M. Valentino, and A. Freitas. 2024. Semeval-2024 task 2: Safe biomedical natural language inference for clinical trials. *arXiv preprint arXiv:2404.04963*.

Ilker Kesen, Andrea Pedrotti, Mustafa Dogan, Michele Cafagna, Emre Can Acikgoz, Letitia Parcalabescu, Iacer Calixto, Anette Frank, Albert Gatt, Aykut Erdem, and Erkut Erdem. 2024. Vilma: A zero-shot benchmark for linguistic and temporal grounding in video-language models. In *International Conference on Learning Representations (ICLR)*.

V. Kougia, A. Sedova, A. Stephan, K. Zaporojets, and B. Roth. 2024. Analysing zero-shot temporal relation extraction on clinical notes using temporal consistency. *arXiv preprint arXiv:2406.11486*.

Joseph Lee, Tianqi Shang, Jae Young Baik, Duy Duong-Tran, Shu Yang, Lingyao Li, and Li Shen. 2025. Investigating llms in clinical triage: Promising capabilities, persistent intersectional biases. *arXiv preprint arXiv:2504.16273*.

Wendy L Macias-Konstantopoulos, Kimberly A Collins, Rosemarie Diaz, Herbert C Duber, Courtney D Edwards, Antony P Hsu, Megan L Ranney, Ralph J Riviello, Zachary S Wettstein, and Carolyn J Sachs. 2023. Race, healthcare, and health disparities: A critical review and recommendations for advancing health equity. *West J Emerg Med*, 24(5):906–918.

D. MacPhail, D. Harbecke, L. Raithel, and S. Möller. 2024. Evaluating the robustness of adverse drug event classification models using templates. *arXiv preprint arXiv:2407.02432*.

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, and 1 others. 2023. Towards accurate differential diagnosis with large language models. *arXiv preprint arXiv:2312.00164*.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

OpenAI(gpt-4.1-mini). 2025. Gpt-4.1 mini. https://platform.openai.com/docs/models/gpt-4.1-mini. Large language model.

OpenMeditron. 2025. Meditron-3-phi4-14b. https://huggingface.co/OpenMeditron/Meditron3-Phi4-14B. Accessed: 2025-05-13.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, Dublin, Ireland. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*.

Tom Röhr, Alexei Figueroa, Jens-Michalis Papaioannou, Conor Fallon, Keno Bressem, Wolfgang Nejdl, and Alexander Löser. 2024. Revisiting clinical outcome prediction for mimic-iv. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 208–217.

Royal College of Physicians. 2017. National early warning score (news) 2. https://www.rcplondon.ac.uk/projects/outputs/national-early-warning-score-news-2.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.

Society of Critical Care Medicine. 2015. Targeted temperature management after cardiac arrest. https://www.sccm.org/blog/concise-critical-appraisal-temperature-management-after-cardiac-arrest.

Society of Critical Care Medicine. 2021. Surviving sepsis campaign: International guidelines for the management of sepsis and septic shock 2021. https://www.sccm.org/SurvivingSepsisCampaign/Guidelines/Adult-Patients.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

B. Van Aken, S. Herrmann, and A. Löser. 2021a. What do you see in this patient? behavioral testing of clinical nlp models. *arXiv preprint arXiv:2111.15512*.

Betty Van Aken, Jens-Michalis Papaioannou, Manuel Mayrdorfer, Klemens Budde, Felix A Gers, and Alexander Loeser. 2021b. Clinical outcome prediction from admission notes using self-supervised knowledge integration. *arXiv preprint arXiv:2102.04110*.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, and 1 others. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142.

World Health Organization. 2016. Oxygen therapy for children: A manual for health workers. https://www.who.int/publications/i/item/9789241549554.

Jie Xu, Lu Lu, Sen Yang, Bilin Liang, Xinwei Peng, Jiali Pang, Jinru Ding, Xiaoming Shi, Lingrui Yang, Huan Song, and 1 others. 2023. Medgpteval: A dataset and benchmark to evaluate responses of large language models in medicine. *arXiv preprint arXiv:2305.07340*.

Zonghai Yao, Zihao Zhang, Chaolong Tang, Xingyu Bian, Youxia Zhao, Zhichao Yang, Junda Wang, Huixue Zhou, Won Seok Jang, Feiyun Ouyang, and 1 others. 2024. Medqa-cs: Benchmarking large language models clinical skills using an ai-sce framework. *arXiv preprint arXiv:2410.01553*.

T. Zack, E. Lehman, M. Suzgun, J. A. Rodriguez, L. A. Celi, J. Gichoya, and E. Alsentzer. 2024. Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 6(1):e12–e22.

Y. Zhao, H. Wang, Y. Liu, W. Suhuang, X. Wu, and Y. Zheng. 2024. Can llms replace clinical doctors? exploring bias in disease diagnosis by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13914–13935.

A Appendix

A.1 Raw Clinical Note Examples

Correct Discharge Note (Raw).

PRESENT ILLNESS: The patient is a ___ year-old female with a history of NSCLC (stage IV) who presents with shortness of breath. (. . .)
MEDICAL HISTORY: CAD s/p MI ___, s/p CABG ___, Hypertension, Dyslipidemia, CVA: small left posterior frontal infarct in ___, Macular Degeneration, NSCLC-stage IV. (. . .)
MEDICATION ON ADMISSION: amlodipine 5 mg, atorvastatin [Lipitor] 80 mg, calcitriol 0.25 mcg, clopidogrel [Plavix] 75 mg, folic acid 1 mg, furosemide 40 mg. (. . .)
ALLERGIES: Codeine
PHYSICAL EXAM: On Admission: Vitals: T: 96.9, BP: 118/51, **HR: 94** , RR: 18, O2Sat: 94% on 5L with face tent.
FAMILY HISTORY: Father died of CAD; mother had stomach cancer and osteosarcoma.
SOCIAL HISTORY: ___.

Counterfactual Discharge Note (Raw). Identical to the correct note, except for the heart rate in **PHYSICAL EXAM**, which is replaced by **HR: 120** instead of **HR: 94**. We do not reproduce the entire note to avoid clutter.

A.2 Template-based clinical note examples

Correct Discharge Note (Template). The original discharge note example shown in template-based format:

Age: 67
Gender: F
Ethnicity: White
Vitals:
Heart Rate: 94
Blood Pressure: 118/51
Respiration Rate: 18
Temperature: 96.9
Oxygen Saturation: 94%

Counterfactual Discharge Note (Template).

Age: 67
Gender: F
Ethnicity: White
Vitals:
Heart Rate: 120
Blood Pressure: 118/51
Respiration Rate: 18
Temperature: 96.9
Oxygen Saturation: 94%

A.3 Extraction of vital signs Specification

To extract vital signs from unstructured clinical notes, we use a few-shot prompt applied to the `PHYSICAL EXAM` section of each note. The model is instructed to extract the vital signs when present.

Prompt template:

You are a clinical information extraction assistant. Your task is to extract the vitals from the PHYSICAL EXAM section of a clinical note.

- Temperature
- Heart Rate (or Pulse)
- Blood Pressure
- Respiration Rate
- Oxygen Saturation

The vitals text may present these values in various formats.

Example Input 1:

```
{"subject_id":"12345","hamd_id":"abcde","sectio
↪ n":"PHYSICAL EXAM","content":"Vitals: T 97.7,
↪ HR 110, BP 99/62, RR 25, O2 99%"}
```

Example Output 1:

```
{"subject_id":"12345","hamd_id":"abcde","vitals
↪ ":{"temperature":"97.7","heart_rate":"110",
↪ "blood_pressure":"99/62","respiration_rate"
↪ :"25","oxygen_saturation":"99%"}}
```

Example Input 2:

```
{"subject_id":"12347","hamd_id":"abcde","sectio
↪ n":"PHYSICAL EXAM","content":"T 98.2, P 117,
↪ O2 97%"}
```

Example Output 2:

```
{"subject_id":"12347","hamd_id":"abcde","vitals
↪ ":{"temperature":"98.2","heart_rate":"117",
↪ "blood_pressure":"","respiration_rate":"","
↪ oxygen_saturation":"97%"}}
```

IMPORTANT: Return ONLY the JSON object and nothing else.

A.4 Clinical guideline-based severity scales

The goal of the severity scale is to create a standardized categorical representation of vital signs that enables aggregation, comparison, and interpretation across variables and experiments. Vital signs differ substantially in their numerical ranges, units, and clinical interpretation. To support a unified counterfactual evaluation framework, we map each continuous vital sign to a shared set of severity categories (`very low`, `low`, `normal`, `high`, `very high`). This standardization allows results to be aggregated across vital signs while preserving clinically meaningful magnitude and directionality of change. We manually derive severity ranges for each vital sign from established clinical guidelines (Infectious Diseases Society of America, 2003; Royal College of Physicians, 2017; Society of Critical Care Medicine, 2015; World Health Organization, 2016; Society of Critical Care Medicine, 2021; American Heart Association, 2024). These variables are widely used in early warning, with ranges chosen to reflect clinically interpretable transitions between physiological states rather than precise diagnostic cutoffs. The resulting categories therefore represent suitable severity levels for the understanding of behavioral analyses.

Variable	Value	Missing (%)
Sex (M / F)	55.2% / 44.8%	0%
Race	White: 69.1% Other/Unknown: 13.2% Black: 10.2% Asian/Pacific: 4.0% Hispanic/Latino: 3.5%	0%
Age (years)	63.64 ± 16.85	0%
Temperature (°F)	97.10 ± 7.69	27%
Heart rate (bpm)	83.86 ± 20.47	5.7%
Respiration rate (bpm)	18.93 ± 5.38	12%
Oxygen saturation (%)	96.99 ± 3.49	5.1%
Systolic BP (mmHg)	128.90 ± 24.15	4.3%
Diastolic BP (mmHg)	71.03 ± 15.46	4.3%

Table 4: Cohort-level statistics used for counterfactual testing, including demographics, vitals, and percentage of missing values.

A.5 Full Population Statistics

Table 4 presents cohort statistics, including demographics, vitals, and missing data percentages.

A.6 Zero-shot prompts

We evaluate models in a zero-shot setting on two downstream tasks: ICU length-of-stay (LoS) prediction and mortality prediction.

ICU length-of-stay (LoS) prediction. Models are asked to predict the LoS class based on the patient's admission note.

You are an expert in ICU length-of-stay prediction. Based only on the patient's admission note, predict in which ICU length-of-stay bucket the patient will fall.

We divide ICU length of stay (for stays $\geq$ 24 hours) into 4 groups:

- `[[1]]` Very short stay (24 to 38 hours)

- [[2]] Short stay (39 to 59 hours)
- [[3]] Moderate stay (60 to 112 hours)
- [[4]] Long stay (113 to 657 hours)

Return only the bucket in double brackets: [[1]], [[2]], [[3]], or [[4]].

In-hospital mortality prediction. Models are asked to predict whether the patient will die during the hospital admission.

You are an expert in ICU mortality prediction. Based only on the patient's structured admission note, predict whether the patient will die before ICU discharge.

Return only the target value in double brackets:

- [[0]] for survival
- [[1]] for death

A.7 Results

A.7.1 Model Specifications

Table 5 summarizes the LLMs used, including model type, domain, and background info on training.

A.7.2 Additional results

Here we present results of additional experiments and for all models. We first report experiments for standard machine learning models—logistic regression (LR), XGBoost, random forest (RF)—when trained on template-based data in Table 6. We observe similar performance between LLMs and these machine learning models on both tasks.

In Table 7 we report precision and recall per model, task and note modality.

In Figure 6 we report task-independent experiments with LLMs, including GPT-OSS-120B. GPT-OSS-120B shares the same behavior as other LLMs, but with a higher magnitude.

In Figure 7 we report change probability of mortality as a function of coun- terfactual severity shift, including LLaMA-3.3-Instruct-70B. This model presented a considerably higher effect than other models and with a different behaviour, overreacting to vital sign perturbations, even whenthe severity stayed the same as the original.

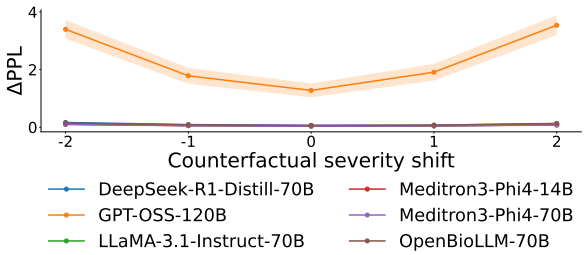


Figure 6: Average per-token ΔPPL across counterfactual severity shifts (raw notes). ΔPPL grows with both increasing and decreasing severity, indicating consistent linguistic sensitivity. GPT-OSS-120B presents a substantially higher effect (2.5 ± 5.1). GPT-4.1-mini is omitted because we could not obtain per-token log probabilities for this model.

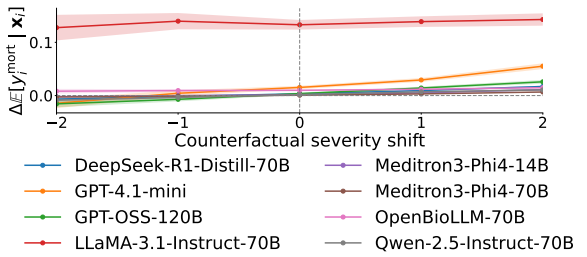


Figure 7: Change of probability of mortality as a function of counterfactual severity shift. Positive severity shifts are expected to increase predicted mortality risk, while negative shifts are expected to decrease it. Most models follow this monotonic trend, indicating clinically aligned responses to vital sign counterfactual perturbations.

LLM	Base LLM	#Params.	Domain	Notes
OpenBioLLM (Ankit Pal, 2024)	Llama-3.3-70B-Instruct	70B	Biomedical	Outperforms GPT-4, Gemini, Meditron, and Med-PaLM-2 on biomedical benchmarks.
Meditron3-Phi4 (OpenMeditron, 2025)	Phi-4	14B	Biomedical	Finetuned version of Phi-4 on medical corpora.
Phi 4 (Abdin et al., 2024)	Phi-4	14B	General-purpose	Trained for efficient language understanding and reasoning.
Llama 3.3 Instruct (Grattafiori et al., 2024)	Llama-3.3-70B-Instruct	70B	General-purpose	Instruction-tuned. Strong performance on general reasoning and text-based tasks.
DeepSeek R1 Distill (Guo et al., 2025)	Llama-3.3-70B-Instruct	70B	General-purpose (reasoning-focused)	Distilled from DeepSeek-R1 using Llama 3.3-70B-Instruct; optimized for multi-step reasoning.
Qwen 2.5 Instruct (Team, 2024)	Qwen-2.5-72B-Instruct	72B	General-purpose	Strong open-source baseline, with competitive performance relative to other large instruction-tuned LLMs.
GPT OSS (OpenAI, 2025)	GPT-OSS-120B	120B	General-purpose	Open-weight GPT model. Large-scale alternative to proprietary LLMs, enabling comparison between open and closed models.
GPT 4.1 mini (OpenAI(gpt-4.1-mini), 2025)	GPT-4.1-mini	N/A	General-purpose	Proprietary model serving as a lightweight reference for commercial systems, included to contrast open-weight models.

Table 5: List of large language models (LLMs) evaluated in this study, including architecture, domain specialization, and training context.

	Mort				LoS			
Model	**Acc.**	**F1**	**Prec.**	**Rec.**	**Acc.**	**F1**	**Prec.**	**Rec.**
LR	0.65	0.21	0.13	0.66	–	–	–	–
XGBoost	0.74	0.22	0.14	0.52	0.29	0.29	0.30	0.29
RF	0.93	0.00	0.00	0.00	0.28	0.28	0.28	0.28

Table 6: Baseline performance for mortality (Mort) and length-of-stay (LoS). **LR**: Logistic regression. **RF**: Random forest. LR is not applicable to LoS (shown as –).

Model	Prec	Rec
Mortality – raw notes		
DeepSeek-R1-Distill-70B	0.55	0.69
GPT-4.1-mini	0.56	0.73
GPT-OSS-120B	**0.61**	0.67
LLaMA-3.3-Instruct-70B	0.57	0.73
Meditron3-Phi4-14B	0.54	0.64
OpenBioLLM-70B	0.60	**0.74**
Phi4-14B	0.53	0.58
Qwen-2.5-Instruct-72B	0.56	0.72
Average	0.57	0.69
Mortality – template-based notes		
DeepSeek-R1-Distill-70B	0.54	0.59
GPT-4.1-mini	0.55	0.59
GPT-OSS-120B	0.47	0.50
LLaMA-3.3-Instruct-70B	0.54	0.59
Meditron3-Phi4-14B	0.56	0.60
OpenBioLLM-70B	**0.63**	0.57
Phi4-14B	0.54	**0.63**
Qwen-2.5-Instruct-72B	0.59	0.60
Average	0.55	0.58
LoS – raw		
DeepSeek-R1-Distill-70B	0.26	0.26
GPT-4.1-mini	0.34	0.32
GPT-OSS-120B	0.29	0.27
LLaMA-3.3-Instruct-70B	0.30	**0.31**
Meditron3-Phi4-14B	0.21	0.28
OpenBioLLM-70B	0.30	0.28
Phi4-14B	0.22	0.28
Qwen-2.5-Instruct-72B	**0.36**	0.29
Average	0.29	0.29
LoS – template		
DeepSeek-R1-Distill-70B	0.22	0.26
GPT-4.1-mini	0.25	0.26
GPT-OSS-120B	0.19	**0.26**
LLaMA-3.3-Instruct-70B	**0.25**	0.24
Meditron3-Phi4-14B	0.17	0.24
OpenBioLLM-70B	0.16	0.24
Phi4-14B	0.18	0.23
Qwen-2.5-Instruct-72B	**0.28**	**0.28**
Average	0.21	0.25

Table 7: Precision and recall for each model variant across note types and tasks. Orange: medical LLM; blue: reasoning-oriented; white: general purpose; gray: averages per block.