# DeVisE: Towards the Behavioral Testing of Medical Large Language Models

Anonymous ACL submission


## 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 probe 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 five LLMs, spanning general-purpose and medical variants, under zero-shot and fine-tuned settings. Model behavior is analyzed through (1) input-level sensitivity, capturing how counterfactuals alter likelihoods, and (2) downstream reasoning, measuring their effect on predicted ICU length-of-stay. Zero-shot models exhibit more coherent and clinically aligned reasoning patterns, whereas fine-tuned ones are stabler but less sensitive to meaningful changes. Demographic factors subtly yet consistently shape predictions, emphasizing the need for fairness-aware evaluation in medical LLMs.

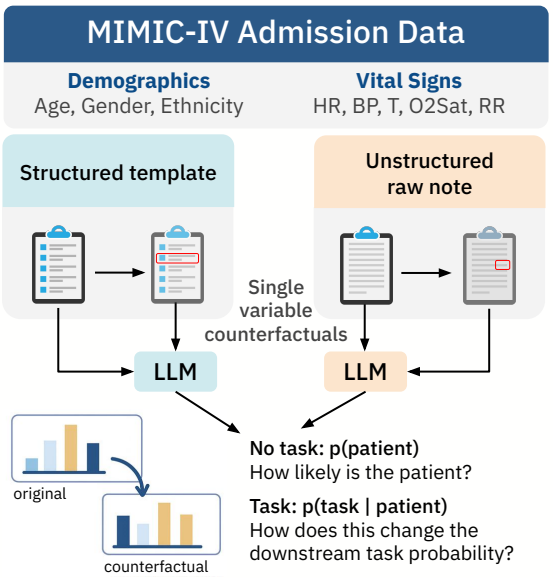




Figure 1: **Overview of DeVisE**. We construct a dataset of 1,000 MIMIC-IV discharge summaries with manually validated single-variable counterfactuals targeting key clinical attributes: **demographics** (age, gender, ethnicity) and **vital signs** (heart, respiration rate, oxygen saturation, temperature, blood pressure). We probe LLMs' sensitivity to input perturbations (e.g., "How does a change in age affect the model's belief in this note?") and downstream effects (e.g., "How does this change affect predicted length-of-stay?").

## 1 Introduction

Large language models (LLMs) are increasingly applied in the medical domain and show strong performance across clinical tasks (McDuff et al., 2023; Singhal et al., 2025; Van Veen et al., 2024; Gu et al., 2021). However, conventional medical benchmarks (Yao et al., 2024; Bakhshandeh, 2023; Xu et al., 2023) primarily rely on coarse metrics like AUROC or F1, offering limited insight into whether models engage in genuine medical reasoning (Van Aken et al., 2021; MacPhail et al., 2024; Jullien et al., 2024) or rely on shortcuts and spurious correlations. Models that excel on these metrics can still fail in settings requiring fine-grained clinical reasoning (Aguiar et al., 2024; Ceballos-Arroyo et al., 2024).

We build on the idea (Fig. 1) that when a clinically meaningful variable (e.g., blood pressure) changes from normal to very high, an LLM should adjust the likelihood of associated outcomes, such as a longer predicted hospital stay. Altering a single attribute may also yield an implausible patient profile, and the model's predictions should reflect that inconsistency.

We present **DeVisE**, a behavioral testing benchmark for clinical NLP based on MIMIC-IV discharge summaries (Johnson et al., 2023). It fo-

cuses on *minimally differing counterfactuals* in core clinical variables: *demographics* (age, gender, ethnicity) and *vital signs* (heart rate, respiration rate, oxygen saturation, and blood pressure). DeVisE includes 1,000 high-quality, manually validated clinical notes. By assessing model behavior without white-box access to a model (Beizer and Wiley, 1996), behavioral testing (BT) enables transparent evaluation of whether predictions align with clinically meaningful reasoning, a key requirement since patient outcomes can depend on subtle variations in only a few variables (Downey et al., 2017; Alghatani et al., 2021; Herasevich et al., 2022).

Because clinical notes are noisy and filled with abbreviations, we also construct synthetic, template-based notes containing only the variables of interest. Analyzing both real and synthetic notes disentangles the effects of linguistic complexity from variable sensitivity (Adnan et al., 2020; Zhang et al., 2020). To our knowledge, prior work has mostly focused on structures templates or fully synthetic inputs (Van Aken et al., 2021; MacPhail et al., 2024; Aguiar et al., 2024; Rajagopal et al., 2021), leaving the behavioral evalution of medical LLMs on realistic text largely unexplored.

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 $1,000$ raw and $1,000$ template-based validated notes, each with corresponding counterfactuals, enabling controlled evaluation of model sensitivity.
- We compare five state-of-the-art LLMs across multiple dimensions: fine-tuned vs. zero-shot, medical vs. general-purpose, reasoning-oriented vs. non-reasoning, and input-only vs. downstream performance (Fig. 1).

This work takes an initial step toward behavioral testing of clinical 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 Related Work

Large language models (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. (2021) 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 assessments with contrast sets to assess 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 (BT)* offers a complementary approach that assesses system behavior without white-box access. BT constructs minimally differing counterfactuals and checks 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).

Building on this paradigm, we introduce **DeVisE**, a behavioral evaluation framework that probes model sensitivity to demographic and vital-sign variations in both raw and template-based clinical notes, bridging clinical reasoning analysis and fairness auditing.

## 3 Methods

**DeVisE** is a behavioral testing framework designed to evaluate two key capabilities of medical LLMs: sensitivity to *demographic variables* and to *vital signs*. Sensitivity to demographics is critical for

detecting bias, while sensitivity to vitals reflects a model's ability to perform numerical reasoning. We test these aspects by constructing counterfactual admission summaries where only one clinical variable is modified at a time.

### 3.1 Clinical variables

**Demographics**. We focus on *gender*, *age*, and *ethnicity*, chosen for their availability in clinical notes and their documented influence on model bias (Van Aken et al., 2021; Celi et al., 2022; Zhao et al., 2024).

**Vital signs**. We consider *heart rate*, *respiration rate*, *oxygen saturation*, *blood pressure*, and *temperature*, all clinically relevant indicators in early detection and prognosis across clinical specialties (Downey et al., 2017; Alghatani et al., 2021; Herasevich et al., 2022).

**Clinical guidelines**. We define clinically meaningful severity classes for each vital sign using evidence-based 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).

### 3.2 Data

**Dataset**. We use discharge summaries from the MIMIC-IV database (Johnson et al., 2023), which document the full course of a hospital stay.

**Raw clinical notes**. Following Röhr et al. (2024) and Van Aken et al. (2021), we extract *only admission-time information* to prevent leakage from post-admission events. That allows us to use these notes to make predictions about the progress of the patient during their stay. We retain the following sections: *chief complaint*, *present illness*, *medical history*, *admission medications*, *allergies*, *physical exam*, *family history*, and *social history*.

**Template-based notes**. We also construct synthetic, noise-free admission notes containing only the variables of interest (demographics and vitals). These allow us to analyze model behavior in a fully controlled setting (see Appendix A.2 for an example).

**Population statistics**. Our dataset includes 1,000 ICU admissions representative of the adult population in MIMIC-IV: 45% female, mean age 64±17 years, and 31% non-white (see Table 6). Missingness for vital signs was low (<6%) except for respiration rate (12%) and temperature (27%).

| Demographic | Classes |
|---|---|
| Age | 18–35: young adults<br>36–55: middle aged adults<br>56–75: older adults<br>76–100: elderly |
| Gender | Female<br>Male |
| Ethnicity | Asian & Pacific<br>Black<br>Hispanic/Latino<br>Other/Unknown<br>White |

Table 1: Demographic categories used to construct single-variables counterfactuals in DeVisE, grouped by clinically relevant age ranges and standard demographic attributes.

### 3.3 Counterfactuals

**Generation**. We generate counterfactuals by modifying one clinical variable at a time—either a demographic attribute or a vital sign—while keeping the rest of the note unchanged. For each variable, we replaced non-missing values to generate counterfactuals: 33,005 (blood pressure), 28,650 (heart rate), 22,110 (temperature), 19,250 (respiration rate), and 13,286 (oxygen saturation).

Five samples per variable/severity class are drawn uniformly (e.g., Tables 1–2), ensuring balanced coverage of severity levels. We expect small numerical perturbations to induce proportionally small changes in model predictions; large shifts under minor changes indicate poor robustness.

Because vital signs in text often diverge from structured data, we extract numerical values from the *physical exam* section using few-shot prompting with LLaMA 3 70B (Grattafiori et al., 2024). We experimented with rule-based extraction via regular expressions but found it less reliable.

**Validation**. We ensure counterfactual consistency through automated and manual checks. Using the GNU diff tool[1] we verify that edits affect only the intended variable, flagging anomalies for human review. Manual validation revealed a 5% error rate in LLM-extracted vitals, while changes in demographics, drawn from structured data, had zero errors. The final test set comprises 1,000 manually verified notes and their associated counterfactuals.

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

| Vital Sign | Classes | Range | Sev. |
|---|---|---|---|
| Heart Rate (bpm) | LTL / Very low | 1–40 | -3 |
| | Low | 41–50 | -1 |
| | Normal | 51–90 | 0 |
| | High | 91–110 | +1 |
| | Very high | 111–130 | +2 |
| | LTH | 131–200 | +3 |
| Blood Pressure (mmHg) | LTL / Very low | 1–70/1–40 | -3 |
| | Low | 71–89 / 41–59 | -1 |
| | Normal | 90–119 / 60–79 | 0 |
| | Elevated | 120–129 / 60–79 | +1 |
| | High | 130–139 / 80–89 | +2 |
| | Very high | 140–179 / 90–119 | +3 |
| | LTH | 80–220/120–140 | +4 |
| Respiration Rate (bpm) | Very low | 1–8 | -2 |
| | Low | 9–11 | -1 |
| | Normal | 12–20 | 0 |
| | High | 21–24 | +1 |
| | Very high | 25–50 | +2 |
| Oxygen Saturation (% $SpO_2$) | LTL | 1–91 | -3 |
| | Very low | 92–93 | -2 |
| | Low | 94–95 | -1 |
| | Normal | 96–100 | 0 |
| Temperature (°F) | LTL | 70.0–82.4 | -3 |
| | Very low | 82.5–89.4 | -2 |
| | Low | 89.5–94.9 | -1 |
| | Normal | 95.0–100.2 | 0 |
| | High | 100.3–103.9 | +1 |
| | LTH | 104.0–110.0 | +3 |

Table 2: Clinically meaningful ranges and severity scores (**Sev.**) used to generate vital-sign counterfactuals. **LTL**: life-threateningly low; **LTH**: life-threateningly high. Severity values follow established clinical guidelines (Section 4).

| Class | Length-of-stay | Reference point |
|---|---|---|
| 1 | $\leq 3$ days | 3 days |
| 2 | $> 3$ and $\leq 7$ days | 7 days |
| 3 | $> 7$ and $\leq 14$ days | 14 days |
| 4 | $> 14$ days | 21 days |

Table 3: Definition of length-of-stay (LOS) classes used in downstream prediction. Reference points correspond to the mean LOS for each class in the training data and are used to compute expected LOS shifts.

## 4 Experimental Setup

We evaluate LLM robustness to minimal edits in demographics and vital signs through two complementary analyses: (1) changes in log-probabilities between original and counterfactual notes, and (2) shifts in predicted length-of-stay (LOS) distributions.

**Length-of-Stay (LOS).** LOS is a challenging but clinically meaningful task, central to hospital resource planning and patient management (McMullan et al., 2004), and is known to correlate with admission vitals (Zebin et al., 2019; Naemi et al., 2021). Using counterfactuals to isolate vitals provides a controlled setting where LOS is an ideal proxy for reasoning quality. Following (Röhr et al., 2024), we define four LOS classes (Table 3) and frame the task as a 4-way classification problem.

**Admission Note Probabilities.** We compute the mean per-token log-probability of each admission note and compare distributions between original and counterfactual versions. Differences in mean and variance indicate a model's sensitivity and *surprise* to small, clinically meaningful perturbations.

**Zero-shot and Fine-tuned LOS.** Models are evaluated both zero-shot and fine-tuned settings for 4-way LOS prediction. During prompting, models are instructed to output the predicted class enclosed in brackets, and token probabilities are used to compute classification metrics. Fine-tuning uses the same MIMIC-IV admission summaries, enabling comparison between general-purpose and medical-domain models to assess whether task alignment improves robustness. Details are provided in Appendix A.3.

**Severity score.** We define a severity scale ranging from -3 to +4, reflecting deviation from the normal range of each vital sign (Table 2). Different vitals map to different severity boundaries; larger absolute values correspond to greater clinical abnormality.[2] The *change in severity* is computed as the difference between the counterfactual and original severity scores. **Example #1:** blood pressure changes from "normal" (0) and "life-threateningly high" (+4), yielding a +4 shift. **Example #2:** temperature changes from "life-threateningly low" (-3) to "high" (+1), yielding -2. These allow us to test whether model outputs vary in clinically consistent directions.

### 4.1 Evaluation Metrics

We compare predicted LOS distributions between original and counterfactual notes using:

- **Jensen–Shannon divergence (JSD)**, measuring the divergence between probability distributions[3]; and

[2] Ranges follow clinical guidelines mentioned in Section 3.3.

[3] $JSD(P||Q) = \frac{1}{2}KL(P||Q) + \frac{1}{2}KL(Q||P)$

- **Expected LOS shift ($\Delta\mathbb{E}_{\text{LOS}}$)**, the difference in expected LOS (in days) between the two distributions, computed using per-class reference points (Table 3).

We further report behavioral consistency metrics:

- **Percentage of correct direction (%Corr Dir)** measures the proportion of cases where $\Delta\mathbb{E}_{\text{LOS}}$ aligns with the direction of severity change (e.g. higher severity → longer stay).
- **Percentage of monotonicity (%Mono)** tracks the proportion of cases where the predicted $\Delta\mathbb{E}_{\text{LOS}}$ follows a consistent upward/downward trend across severity levels. For each of the 9 severity levels shifts (counterfactual severity - original severity), ranging from -4 to +4, we compute the mean predicted $\Delta\mathbb{E}_{\text{LOS}}$. We then evaluate how many consecutive steps follow the expected trend: positive severity values should correspond to positive changes in LOS (longer stays), and negative severity values to negative changes (shorter stays). A step is considered correct if both the sign of the change and the direction of the trend (increasing or decreasing) align with the progression of the severity.
- **Percentage of flips (%Flip)** measures prediction stability in terms of class labels, unlike previous metrics which focus on the expected LOS values. It looks at how often the most probable LOS class changes between the original and counterfactual inputs, indicating whether counterfactual perturbations caused categorical prediction shifts.
- **Distribution of LOS shifts** calculates the proportion of counterfactuals that resulted in increased ($\%\Delta E^{+}$) versus decreased ($1 - \%\Delta E^{+}$) expected LOS to understand whether LLMs favor increased or decreased LOS. We also report the average magnitude of these shifts in both directions.

### 4.2 Large language models (LLMs)

We evaluate five LLMs of varying architectures, domains, and reasoning orientations: OpenBioLLM (Ankit Pal, 2024) (70B) and Meditron3-Phi4 (OpenMeditron, 2025) (14B), fine-tuned on medical data, three general-purpose models, Phi 4 (Abdin et al., 2024) (14B), LLaMA 3 Instruct (Grattafiori et al., 2024) (70B), and DeepSeek R1 Distill (dsR1) (Guo et al., 2025) (70B - reasoning-oriented). We analyze LLMs along three axes: (1) *fine-tuned vs. zero-shot*, (2) *medical-domain vs. general-purpose*, and (3) *reasoning-oriented vs. non-reasoning*. These comparisons reveal how model specialization and reasoning capacity affect robustness, sensitivity, and behavioral consistency. Model specifications are summarized in Table 8. This subset is not intended to be exhaustive, but captures diversity across base architecture, size, domain, and reasoning focus, offering a representative analysis of current LLM behavior. *We do not use proprietary LLMs since the data usage agreement of MIMIC-IV does not allow for it.*[4]

## 5 Results

First, we note that the performance of our fine-tuned LLMs on LOS prediction (in Table 4) aligns with that of state-of-the-art models in the literature (e.g., van Aken et al., 2021).

### 5.1 Vital signs

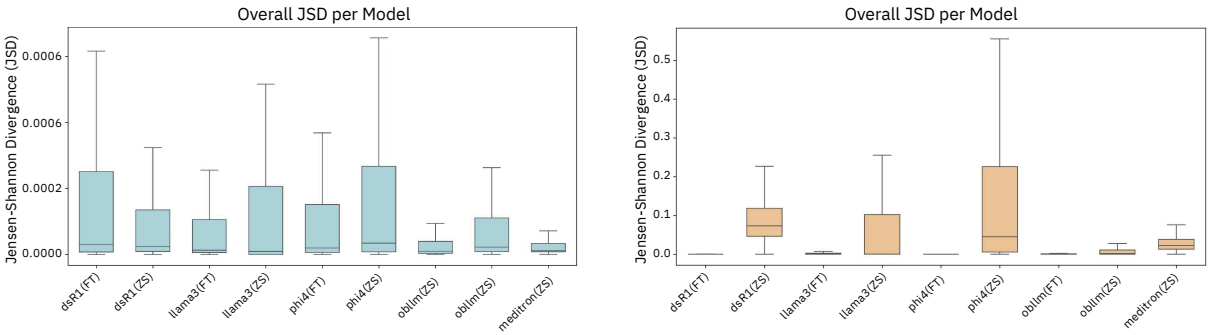
Figure 2 compares distributional shifts (JSD) across models under vital sign perturbations. On raw notes, counterfactual edits cause minimal divergence (Figure 2a), while template-based notes yield substantially higher shifts, especially in zero-shot (ZS) settings (Figure 2b). Fine-tuned (FT) models achieve modestly higher F1 scores than ZS counterparts (Table 4), but ZS models exhibit greater behavioral sensitivity, with elevated $\Delta\mathbb{E}_{\text{LOS}}$, %Flip, and %Mono values. Flip rates in FT models remain below 1% across all conditions. Raw notes produce higher %Mono than templates, indicating more stable directional consistency.

The relationship between perturbation severity and $\Delta\mathbb{E}_{\text{LOS}}$ (Figure 3) is more clinically coherent for raw notes. ZS models show larger changes on templates, but often fail to decrease LOS with decreasing severity, revealing weaker monotonicity. All models tend to predict longer stays after perturbations ($\%\Delta E^{+} > 50$), with the strongest effect in ZS template-based settings (e.g., Meditron: 54% → 96%). This suggests an overall bias toward associating any vital change with worsening outcomes, *regardless of the direction of change*.

Behavioral differences also emerge without an explicit prediction task. Changes in token-level log-likelihoods remain small but consistently scale with severity (Table 5, Figure 4), implying that models are linguistically *surprised* by counterfactuals even when not trained for LOS prediction.

Medical LLMs in the FT setting are more con-

[4] https://physionet.org/about/licenses/physionet-credentialed-health-data-license-150/

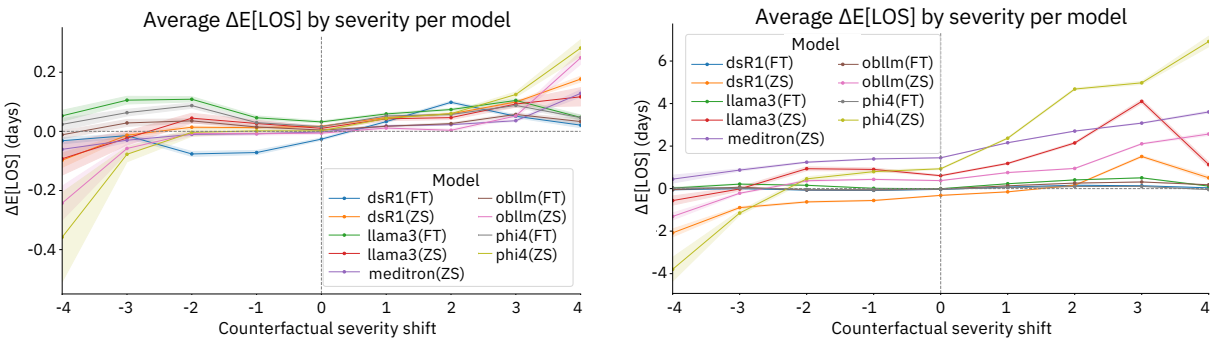



(a) Raw notes (b) Template-based notes

Figure 2: Jensen–Shannon divergence (JSD) per model under vital-sign perturbations. (a) Raw notes show minimal distributional change across models. (b) Template-based notes yield substantially higher JSD, especially for zero-shot (ZS) models compared to fine-tuned (FT) ones. Y-axes differ between panels, and outliers are omitted for clarity.

| Model | Setting | Acc | F1 | % $\Delta E^+$ | Avg $\Delta E^+$ / $\Delta E^-$ | Std $\Delta E^+$ | Std $\Delta E^-$ | Avg JSD | Std JSD | %Corr Dir | %Flip | %Mono | Top 1 Vital |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| dsR1(FT) | raw | **0.723** | 0.220 | 56 | 0.142 / -0.134 | 0.357 | 0.258 | 0.001 | 0.006 | 48 | 0 | 50 | respiration rate |
| llama3(FT) | raw | 0.722 | **0.311** | 63 | 0.130 / -0.057 | 0.356 | 0.098 | 0.001 | 0.006 | 42 | 0 | 33 | blood pressure |
| phi4(FT) | raw | **0.723** | 0.273 | 64 | 0.126 / -0.096 | 0.308 | 0.199 | 0.001 | 0.005 | 41 | 0 | 33 | blood pressure |
| obllm(FT) | raw | **0.723** | 0.240 | 63 | 0.072 / -0.068 | 0.267 | 0.179 | 0.000 | 0.005 | 43 | 0 | 33 | blood pressure |
| meditron(ZS) | raw | 0.332 | 0.214 | 54 | 0.071 / -0.077 | 0.110 | 0.197 | 0.000 | 0.001 | 43 | 4 | **100** | temperature |
| dsR1(ZS) | raw | 0.163 | 0.120 | 65 | 0.093 / -0.048 | 0.152 | 0.099 | 0.000 | 0.002 | **49** | 2 | 67 | temperature |
| llama3(ZS) | raw | 0.116 | 0.119 | 64 | 0.138 / -0.146 | 0.517 | 0.600 | 0.006 | 0.047 | 42 | 2 | 67 | blood pressure |
| phi4(ZS) | raw | 0.080 | 0.082 | 63 | 0.146 / -0.156 | 0.294 | 0.521 | 0.002 | 0.016 | 42 | 2 | 83 | blood pressure |
| obllm(ZS) | raw | 0.119 | 0.118 | 58 | 0.109 / -0.139 | 0.200 | 0.411 | 0.001 | 0.006 | 37 | 2 | 66 | blood pressure |
| dsR1(FT) | template | 0.722 | 0.210 | 63 | 0.167 / -0.148 | 0.175 | 0.185 | 0.000 | 0.001 | **49** | 0 | 33 | respiration rate |
| llama3(FT) | template | 0.722 | 0.210 | 66 | 0.464 / -0.258 | 0.401 | 0.267 | 0.003 | 0.004 | 44 | 0 | 33 | respiration rate |
| phi4(FT) | template | 0.722 | 0.210 | 59 | 0.138 / -0.121 | 0.216 | 0.200 | 0.001 | 0.002 | 44 | 0 | 33 | respiration rate |
| obllm(FT) | template | 0.722 | 0.210 | 67 | 0.271 / -0.194 | 0.227 | 0.201 | 0.001 | 0.002 | 47 | 0 | 50 | oxygen saturation |
| meditron(ZS) | template | 0.327 | 0.184 | **96** | 2.198 / -0.429 | 1.040 | 0.525 | 0.029 | 0.026 | 35 | **73** | 50 | temperature |
| dsR1(ZS) | template | 0.217 | 0.113 | 49 | 1.091 / **-1.185** | 1.318 | 0.856 | 0.087 | 0.059 | 37 | 15 | 83 | respiration rate |
| llama3(ZS) | template | 0.258 | 0.149 | 76 | 1.925 / -0.398 | 3.805 | 1.334 | 0.151 | 0.311 | 42 | 18 | 50 | oxygen saturation |
| phi4(ZS) | template | 0.187 | 0.165 | 76 | **3.679** / -0.985 | 3.534 | 1.452 | **0.158** | 0.228 | 40 | 41 | 67 | oxygen saturation |
| obllm(ZS) | template | 0.179 | 0.098 | 75 | 1.033 / -0.311 | 1.189 | 0.481 | 0.009 | 0.020 | 31 | 15 | 67 | oxygen saturation |

Table 4: Performance and behavioral metrics for each model variant across note types. FT = fine-tuned; ZS = zero-shot. Orange = medical LLM; gray = reasoning-oriented, white = general purpose. Highest values are bold, and lowest are underlined. Accuracy differences among FT models are minor, likely reflecting LOS class imbalance (Appendix A.6).




(a) Raw notes (b) Template-based notes

Figure 3: Average change in expected LOS $\Delta\mathbb{E}_{\text{LOS}}$ across severity levels for each model. (a) Raw notes produce smoother, clinically consistent trends; (b) Template-based notes amplify magnitude.

servative, showing lower JSD and $\Delta\mathbb{E}_{\text{LOS}}$, yet comparable in terms of accuracy and directional correctness to general-purpose models. Conversely, medical ZS models attain higher accuracy and F1 but exhibit larger flip rates. dsR1 achieves the highest %Corr Dir across most settings and remains notably stable in ZS raw notes.

Figure 5 decomposes JSD by vital sign. In raw

| Model | Avg $\Delta$ loglik $\pm$ Std |
|---|---|
| dsR1 | $-0.0015 \pm 0.0083$ |
| llama3 | $-0.0019 \pm 0.0116$ |
| meditron | $-0.0012 \pm 0.0128$ |
| obllm | $-0.0008 \pm 0.0086$ |
| phi4 | $-0.0011 \pm 0.0113$ |

Table 5: Mean change in per-token log-likelihood (±std). under counterfactual perturbations. Larger magnitudes indicate stronger linguistic sensitivity to variable edits.

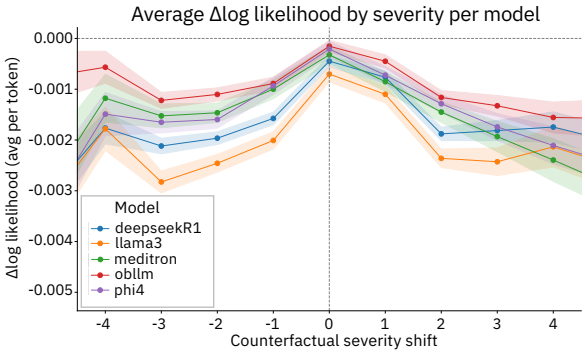



Figure 4: Average per-token $\Delta$log-likelihood across severity shifts (raw notes). Log-likelihood changes grow with both increasing and decreasing severity, indicating consistent linguistic sensitivity.

notes, blood pressure dominates model sensitivity; in templates, respiration rate is most influential. Some models, like dsR1(FT), maintain consistent vital focus, whereas others (e.g. phi4, obllm) shift across FT/ZS settings.

### 5.2 Demographic variables

Demographic perturbations also yield consistent, statistically significant effects, particularly in ZS setting with template-based inputs (Figure 6). Among demographic factors, 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 up to 6% in ZS models. While most comparisons are significant, exceptions include gender in obllm(ZS) and the "Other/Unknown" race category in llama3(ZS).

Across models, dsR1(ZS) remains the most stable, showing the lowest $\Delta\mathbb{E}_{\text{LOS}}$ and near-zero flips. In contrast, obllm(ZS) and phi4(ZS) produce the largest shifts for race and gender, respectively. Race-related effects are generally larger than gender effects, and all models predict shorter LOS for female and black patients, aligning with known healthcare disparities (Macias-Konstantopoulos et al., 2023).

In FT settings, models become largely insensitive to demographic changes: $\Delta\mathbb{E}_{\text{LOS}}$ is nearly zero and no flips are observed (Figure 8). This suggests that task-specific fine-tuning may suppress both desirable and undesirable demographic sensitivity, highlighting the need for fairness-aware objectives.

Detailed disaggregated results by model and demographic class are presented in Tables 11 and 12 in Appendix A.7.4.

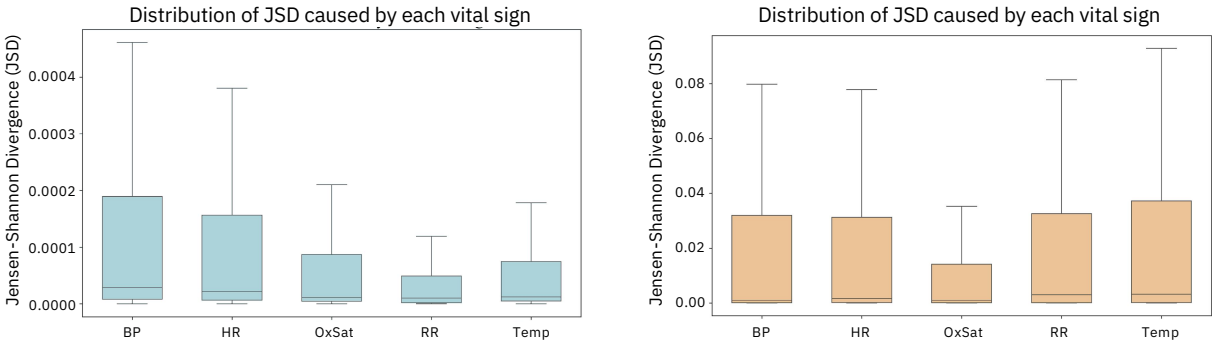
## 6 Discussion

**Fine-tuned vs Zero-shot**. Despite lower accuracy, zero-shot (ZS) models exhibit clearer behavioral patterns across both JSD and LOS shifts (Table 4). This suggests that, without task-specific tuning, they may rely more on generalizable reasoning patterns, e.g., associating low oxygen saturation or extreme blood pressure with longer LOS. In contrast, fine-tuned models are more conservative, often underreacting to clinically significant changes. This may reflect overfitting to skewed LOS label distributions (e.g., bias toward bucket 1), resulting in reduced sensitivity to causal variation. These findings challenge the assumption that fine-tuning necessarily improves robustness, showing that it may instead reinforce dataset-specific heuristic behaviors.

**Medical vs General-purpose**. Domain-specific LLMs, in both ZS and FT modes, reacted less strongly to counterfactuals than general-purpose models (lower JSD and $\Delta\mathbb{E}_{\text{LOS}}$ values; Tables 9 and 10), possibly due to more conservative priors learned from biomedical corpora. However, their higher %Mono indicates closer alignment with clinical expectations. Interestingly, ZS medical models outperform general ones in accuracy despite showing more class flips, suggesting stronger domain grounding even without task-specific tuning.

**Reasoning vs Non-reasoning**. DeepSeek demonstrated a balanced profile of robustness and sensitivity, particularly its fine-tuned variant, achieving high %Corr Dir with minimal flips, especially on raw notes (Fig. 3). Its reasoning-oriented pretraining appears to stabilize predictions without suppressing clinically meaningful variability, highlighting the promise of reasoning-focused objectives for medical LLM reliability.

**Task vs No-task**. Even without explicit supervision, models respond systematically to increasing




(a) Raw notes

(b) Template-based notes

Figure 5: JSD distribution per vital sign across all models. (a) Raw notes show modest divergence dominated by blood pressure; (b) Template-based notes amplify variance, with respiration rate and temperature driving the largest shifts. Outliers omitted for clarity.

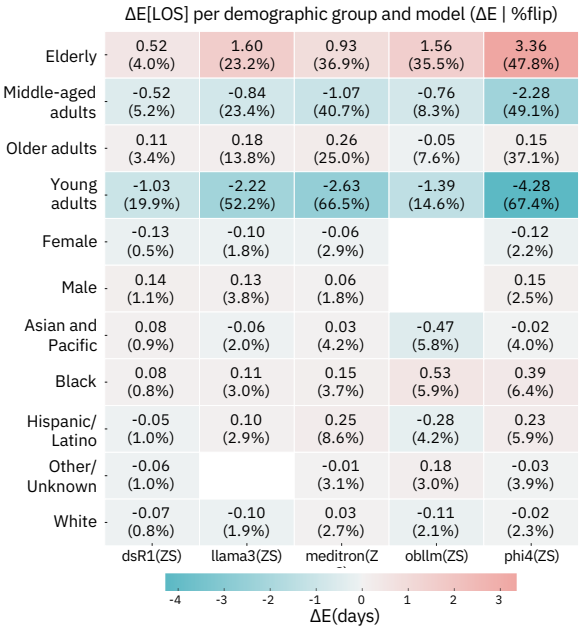



Figure 6: Mean $\Delta\mathbb{E}_{\text{LOS}}$ and flip rates (%) by demographic class for zero-shot models. Age produces the largest and most significant shifts, while race and gender effects are smaller but consistent. Blank cells denote non-significant comparisons.

counterfactual severity. Log-likelihood shifts form a pyramid-like pattern across severity levels (Fig. 4) except for severity -3 to -4, 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**. Template-based notes resulted in more pronounced yet, in general, noisier and less coherent LOS predictions (Fig.3) seen by a lower %Mono and decreased correct direction of LOS predictions. The absence of contextual information may reduce the model's ability to reason about changes. In contrast, raw notes helped models contextualize vital sign changes more effectively, though the importance of vitals was diluted by other information. Variability across both formats (see Figure 3 and Tables 4, 5) highlights ongoing uncertainty in model behavior.

**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. 6) raise fairness concerns. Consistent underprediction of LOS for women and overprediction for black patients across all models suggest that pretraining biases persist even in ZS models, reinforcing the need for demographic auditing in clinical LLM evaluation.

## 7 Conclusion

Our results show the usefulness of examining how medical LLMs respond to clinically-informed counterfactual input changes to understand LLM behavior and reasoning patterns. Our findings reveal that zero-shot models demonstrate coherent reasoning patterns, whereas fine-tuned models are more conservative and show more stable outputs, sometimes even being insensitive to clinically significant changes in the input. Demographic variables affect predictions in a subtle but consistent way, underscoring the need for fairness-aware evaluations. Future research should include fine-tuning strategies that help maintain reasoning capacity and improve calibration and fairness, as well as developing and implementing diagnostic tools to detect and mitigate model bias in real-world clinical settings.

## Limitations

In the following paragraphs, we discuss some limitations of our experimental setup.

**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.

**Severity scales and clinical validation**. We used different clinical guidelines to create severity scores for different vital sign variables. We tried as much as possible to stick to the nomenclature in the clinical guidelines to emphasise clinical relevance. That means that different vital signs can have severity classes with different severity scores, e.g., "high" blood pressure is +2, whereas "high" temperature is +1. We note that this does not affect our analysis, however, since we have conducted only single-variable counterfactuals. We believe future work should look into employing clinicians from different clinical specialties to align the severity classes and scores for different clinical variables.

**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.

## 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*.

Kiran Adnan, Rehan Akbar, Siak Wang Khor, and Adnan Bin Amanat Ali. 2020. Role and challenges of unstructured big data in healthcare. *Data management, analytics and innovation*, pages 301–323.

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.

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

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*.

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 EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and 1 others. 2023. 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*.

R McMullan, B Silke, K Bennett, and S Callachand. 2004. Resource utilisation, length of hospital stay, and pattern of investigation during acute medical hospital admission. *Postgraduate medical journal*, 80(939):23–26.

Amin Naemi, Thomas Schmidt, Marjan Mansourvar, Ali Ebrahimi, and Uffe Kock Wiil. 2021. Prediction of length of stay using vital signs at the admission time in emergency departments. In *Innovation in Medicine and Healthcare: Proceedings of 9th KES-InMed 2021*, pages 143–153. Springer.

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.

Dheeraj Rajagopal, Vivek Khetan, Bogdan Sacaleanu, Anatole Gershman, Andrew Fano, and Eduard Hovy. 2021. Template filling for controllable commonsense reasoning. *arXiv preprint arXiv:2111.00539*.

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/getattachment/Disaster/Disaster-Management/Temperature-Management-After-Cardiac-Arrest/Targeted-Temperature-Management-After-Cardiac-Arrest.pdf.

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.

B. Van Aken, S. Herrmann, and A. Löser. 2021. 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 Gers, and Alexander Loeser. 2021. Clinical outcome prediction from admission notes using self-supervised knowledge integration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 881–893, Online. Association for Computational Linguistics.

Betty Van Aken, Jens-Michalis Papaioannou, Manuel Mayrdorfer, Klemens Budde, Felix A Gers, and Alexander Loeser. 2021. 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.

Tahmina Zebin, Shahadate Rezvy, and Thierry J Chaussalet. 2019. A deep learning approach for length of stay prediction in clinical settings from medical records. In *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–5. IEEE.

Dongdong Zhang, Changchang Yin, Jucheng Zeng, Xiaohui Yuan, and Ping Zhang. 2020. Combining structured and unstructured data for predictive models: a deep learning approach. *BMC medical informatics and decision making*, 20(1):280.

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 Fine-tuning Specification

Only admission-time sections were retained to avoid leakage of future information. Fine-tuning was conducted with the LLaMA Factory framework using LoRA adapters, trained over 3 epochs (learning rate: 5e-5, batch size: 2, gradient accumulation: 4). Evaluation was performed every 10 steps using a held-out validation set.

## A.4 Evaluation Metrics Aggregation Details

**Jensen–Shannon divergence (JSD)** and the **Expected LOS shift** ($\Delta\mathbb{E}_{\text{LOS}}$) are aggregated per admission (by `hadm_id`), then averaged across counterfactuals for each variable to ensure comparability. We further group results by model family (e.g., fine-tuned vs. zero-shot) and perturbation severity.

## A.5 Full Population Statistics

Table 6 presents cohort statistics, including demographics, vitals, and missing data percentages.

## A.6 LOS class representation

We provide additional results about LOS prediction in Table 7.

## A.7 Results

### A.7.1 Model Specifications

Table 8 summarizes the LLMs used, including model type, domain, and background info on training.

| Variable | Value | Missing (%) |
|---|---|---|
| Sex (M / F) | 55.2% / 44.8% | 0% |
| Race | White: 69.1%<br>Other/Unknown: 13.2%<br>Black: 10.2%<br>Asian/Pacific: 4.0%<br>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 6: Cohort-level statistics used for counterfactual testing, including demographics, vitals, and percentage of missing values.

| Bucket | # Cases |
|---|---|
| 1 | 30674 |
| 2 | 10028 |
| 3 | 3872 |
| 4 | 2177 |

Table 7: LOS class distribution in the training set. There is an over-representation of shorter stays making it harder for the FT models to predict longer stay buckets.

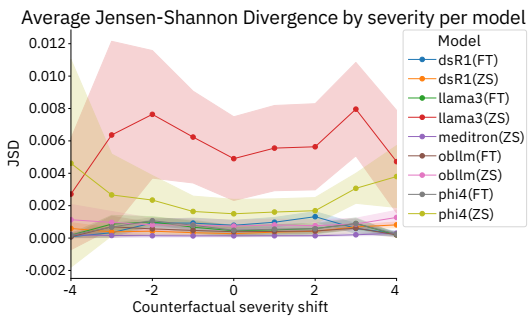



Figure 7: Jensen-Shannon divergence (JSD) across counterfactual severity levels for each model on raw clinical notes. Higher severity generally results in larger distributional shifts.

### A.7.2 Model Group Comparisons

Summary statistics and pairwise significance tests across model groups and settings (FT vs. ZS; template vs. raw) are available in Tables 9 and 10.

### A.7.3 JSD by Severity (Raw Notes)

JSD sensitivity across severity levels per model is visualized in Figure 7.

| LLM | Base | Parameters | Domain | Nickname | Notes |
|---|---|---|---|---|---|
| OpenBioLLM (Ankit Pal, 2024) | Llama-3.3-70B-Instruct | 70B | Biomedical | obllm | Outperforms GPT-4, Gemini, Meditron, and Med-PaLM-2 on biomedical benchmarks. |
| Meditron3-Phi4 (OpenMeditron, 2025) | Phi-4 | 14B | Biomedical | meditron | Finetuned version of Phi-4 on medical corpora. |
| Phi 4 (Abdin et al., 2024) | Phi-4 | 14B | General-purpose | phi4 | Trained for efficient language understanding and reasoning. |
| Llama 3 Instruct (Grattafiori et al., 2024) | Llama-3.3-70B-Instruct | 70B | General-purpose | llama3 | 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) | dsR1 | Distilled from DeepSeek-R1 using Llama 3.3-70B-Instruct; optimized for multi-step reasoning. |

Table 8: List of large language models (LLMs) evaluated in this study, including architecture, domain specialization, and training context.

| Model | Setting | Acc | F1 | $\%\Delta E^+$ | Avg $\Delta E^+$ / $\Delta E^-$ | Std $\Delta E^+$ | Std $\Delta E^-$ | Avg JSD | Std JSD | %Corr Dir | %Mono |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Fine-tuned | raw | 0.723 | 0.261 | 61 | 0.117 / -0.088 | 0.357 | 0.258 | 0.001 | 0.004 | 44 | 63 |
| Zero-shot | raw | 0.162 | 0.131 | 61 | 0.112 / -0.113 | 0.517 | 0.600 | 0.002 | 0.011 | 43 | 88 |
| Medical(FT) | raw | 0.723 | 0.240 | 63 | 0.072 / -0.068 | 0.267 | 0.179 | 0.001 | 0.005 | 43 | 63 |
| General(FT) | raw | 0.723 | 0.292 | 64 | 0.128 / -0.076 | 0.356 | 0.098 | 0.001 | 0.004 | 42 | 63 |
| Reasoning(FT) | raw | 0.723 | 0.220 | 56 | 0.142 / -0.134 | 0.357 | 0.258 | 0.001 | 0.005 | 48 | 63 |
| Medical(ZS) | raw | 0.226 | 0.166 | 56 | 0.090 / -0.108 | 0.110 | 0.197 | 0.001 | 0.003 | 40 | 88 |
| General(ZS) | raw | 0.098 | 0.100 | 64 | 0.142 / -0.151 | 0.294 | 0.521 | 0.004 | 0.025 | 42 | 88 |
| Reasoning(ZS) | raw | 0.163 | 0.120 | 65 | 0.093 / -0.048 | 0.152 | 0.099 | 0.000 | 0.002 | 49 | 88 |
| Fine-tuned | template | 0.722 | 0.210 | 63 | 0.260 / -0.180 | 0.260 | 0.180 | 0.001 | 0.001 | 46 | 63 |
| Zero-shot | template | 0.234 | 0.142 | 74 | 1.985 / -0.662 | 1.985 | 0.662 | 0.094 | 0.071 | 37 | 47 |
| Medical(FT) | template | 0.722 | 0.210 | 67 | 0.271 / -0.194 | 0.271 | 0.194 | 0.001 | 0.001 | 47 | 50 |
| General(FT) | template | 0.722 | 0.210 | 63 | 0.301 / -0.190 | 0.301 | 0.190 | 0.002 | 0.001 | 44 | 75 |
| Reasoning(FT) | template | 0.722 | 0.210 | 63 | 0.167 / -0.148 | 0.167 | 0.148 | 0.000 | 0.001 | 49 | 50 |
| Medical(ZS) | template | 0.253 | 0.141 | 85 | 1.615 / -0.370 | 1.040 | 0.525 | 0.020 | 0.015 | 33 | 50 |
| General(ZS) | template | 0.223 | 0.157 | 76 | 2.802 / -0.691 | 3.534 | 1.452 | 0.168 | 0.160 | 41 | 50 |
| Reasoning(ZS) | template | 0.217 | 0.113 | 49 | 1.091 / -1.185 | 1.091 | 1.185 | 0.092 | 0.048 | 37 | 33 |

Table 9: Aggregate behavioral performance of model groups (fine-tuned (FT) vs. zero-shot (ZS), medical vs. general-purpose vs. reasoning-focused) on raw and template-based notes..

### A.7.4 Demographic variables

Full demographic bias analysis is presented in Tables 11 and 12. Figure 8 visualizes mean LOS shifts across demographic subgroups for FT models. In template-based settings, FT models are highly biased toward LOS bucket 1, yielding 0% flip rates across all demographics.

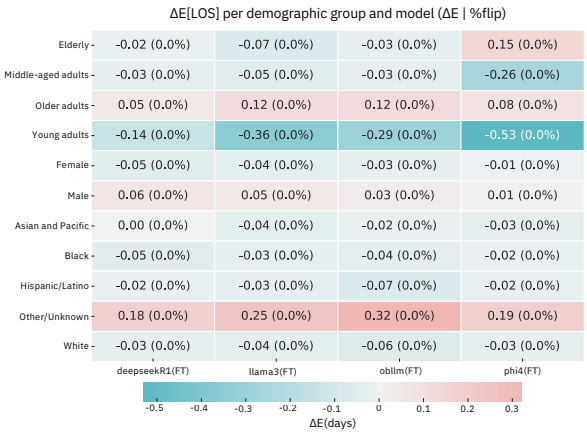



Figure 8: Mean change in expected length-of-stay ($\Delta\mathbb{E}_{\text{LOS}}$) across demographic subgroups for fine-tuned models. All models consistently predict shorter LOS for females and Black patients.

| Comparison | ZS/FT? | Setting | p-value | Winner |
|---|---|---|---|---|
| FT vs ZS (overall) | — | Structured | $< 10^{-10}$ | ZS |
| FT vs ZS (overall) | — | Unstructured | $< 10^{-10}$ | ZS |
| Medical vs. General | FT | Structured | $< 10^{-10}$ | General |
| Medical vs. Reasoning | FT | Structured | $< 10^{-10}$ | Medical |
| General vs. Reasoning | FT | Structured | $< 10^{-10}$ | General |
| Medical vs. General | FT | Unstructured | $10^{-6}$ | General |
| Medical vs. Reasoning | FT | Unstructured | $< 10^{-10}$ | Reasoning |
| General vs. Reasoning | FT | Unstructured | $< 10^{-10}$ | Reasoning |
| Medical vs. General | ZS | Structured | $< 10^{-10}$ | General |
| Medical vs. Reasoning | ZS | Structured | $< 10^{-10}$ | Reasoning |
| General vs. Reasoning | ZS | Structured | $< 10^{-10}$ | General |
| Medical vs. General | ZS | Unstructured | $< 10^{-10}$ | General |
| Medical vs. Reasoning | ZS | Unstructured | 0.059 | Similar |
| General vs. Reasoning | ZS | Unstructured | $< 10^{-10}$ | General |

Table 10: Paired t-test comparisons of average JSD values between model groups. Significance is assessed across note formats and training conditions.The "Winner" column indicates the group with significantly higher JSD or "Similar" when no significant difference is found.

| Model | Variable | Class | Mean ΔE | Std ΔE | p-value | % Class Flip |
|---|---|---|---|---|---|---|
| dsR1(FT) | age | young adults | -0.14 | 0.11 | $< 10^{-10}$ | 0% |
| dsR1(FT) | age | middle aged adults | -0.03 | 0.09 | $< 10^{-10}$ | 0% |
| dsR1(FT) | age | older adults | 0.05 | 0.08 | $< 10^{-10}$ | 0% |
| dsR1(FT) | age | elderly | -0.02 | 0.10 | $< 10^{-10}$ | 0% |
| dsR1(FT) | gender | F | -0.05 | 0.04 | $< 10^{-10}$ | 0% |
| dsR1(FT) | gender | M | 0.06 | 0.04 | $< 10^{-10}$ | 0% |
| dsR1(FT) | race | Asian & Pacific | 0.00 | 0.08 | 0.000 | 0% |
| dsR1(FT) | race | Black | -0.05 | 0.08 | $< 10^{-10}$ | 0% |
| dsR1(FT) | race | Hispanic/Latino | -0.02 | 0.08 | $< 10^{-10}$ | 0% |
| dsR1(FT) | race | Other/Unknown | 0.18 | 0.09 | $< 10^{-10}$ | 0% |
| dsR1(FT) | race | White | -0.03 | 0.08 | $< 10^{-10}$ | 0% |
| dsR1(ZS) | age | young adults | -1.03 | 0.56 | $< 10^{-10}$ | 20% |
| dsR1(ZS) | age | middle aged adults | -0.52 | 0.52 | $< 10^{-10}$ | 5% |
| dsR1(ZS) | age | older adults | 0.11 | 0.55 | $< 10^{-10}$ | 3% |
| dsR1(ZS) | age | elderly | 0.52 | 0.55 | $< 10^{-10}$ | 4% |
| dsR1(ZS) | gender | F | -0.13 | 0.07 | $< 10^{-10}$ | 1% |
| dsR1(ZS) | gender | M | 0.14 | 0.07 | $< 10^{-10}$ | 1% |
| dsR1(ZS) | race | Asian & Pacific | 0.08 | 0.14 | $< 10^{-10}$ | 1% |
| dsR1(ZS) | race | Black | 0.08 | 0.12 | $< 10^{-10}$ | 1% |
| dsR1(ZS) | race | Hispanic/Latino | -0.05 | 0.15 | $< 10^{-10}$ | 1% |
| dsR1(ZS) | race | Other/Unknown | -0.06 | 0.17 | $< 10^{-10}$ | 1% |
| dsR1(ZS) | race | White | -0.07 | 0.11 | $< 10^{-10}$ | 1% |
| llama3(FT) | age | young adults | -0.36 | 0.28 | $< 10^{-10}$ | 0% |
| llama3(FT) | age | middle aged adults | -0.05 | 0.25 | $< 10^{-10}$ | 0% |
| llama3(FT) | age | older adults | 0.12 | 0.27 | $< 10^{-10}$ | 0% |
| llama3(FT) | age | elderly | -0.07 | 0.33 | 0.000 | 0% |
| llama3(FT) | gender | F | -0.04 | 0.03 | $< 10^{-10}$ | 0% |
| llama3(FT) | gender | M | 0.05 | 0.03 | $< 10^{-10}$ | 0% |
| llama3(FT) | race | Asian & Pacific | -0.04 | 0.11 | $< 10^{-10}$ | 0% |
| llama3(FT) | race | Black | -0.03 | 0.11 | $< 10^{-10}$ | 0% |
| llama3(FT) | race | Hispanic/Latino | -0.03 | 0.11 | $< 10^{-10}$ | 0% |
| llama3(FT) | race | Other/Unknown | 0.25 | 0.11 | $< 10^{-10}$ | 0% |
| llama3(FT) | race | White | -0.04 | 0.11 | $< 10^{-10}$ | 0% |
| llama3(ZS) | age | young adults | -2.22 | 2.10 | $< 10^{-10}$ | 52% |
| llama3(ZS) | age | middle aged adults | -0.84 | 1.96 | $< 10^{-10}$ | 23% |
| llama3(ZS) | age | older adults | 0.18 | 1.69 | $< 10^{-10}$ | 14% |
| llama3(ZS) | age | elderly | 1.60 | 3.08 | $< 10^{-10}$ | 23% |
| llama3(ZS) | gender | F | -0.10 | 0.32 | $< 10^{-10}$ | 2% |
| llama3(ZS) | gender | M | 0.13 | 0.35 | $< 10^{-10}$ | 4% |
| llama3(ZS) | race | Asian & Pacific | -0.06 | 0.35 | $< 10^{-10}$ | 2% |
| llama3(ZS) | race | Black | 0.11 | 0.41 | $< 10^{-10}$ | 3% |
| llama3(ZS) | race | Hispanic/Latino | 0.10 | 0.42 | $< 10^{-10}$ | 3% |
| llama3(ZS) | race | Other/Unknown | 0.00 | 0.37 | 0.386 | 3% |
| llama3(ZS) | race | White | -0.10 | 0.39 | $< 10^{-10}$ | 2% |
| phi4(FT) | age | young adults | -0.53 | 0.28 | $< 10^{-10}$ | 0% |
| phi4(FT) | age | middle aged adults | -0.26 | 0.29 | $< 10^{-10}$ | 0% |
| phi4(FT) | age | older adults | 0.08 | 0.27 | $< 10^{-10}$ | 0% |
| phi4(FT) | age | elderly | 0.15 | 0.31 | $< 10^{-10}$ | 0% |
| phi4(FT) | gender | F | -0.01 | 0.02 | $< 10^{-10}$ | 0% |
| phi4(FT) | gender | M | 0.01 | 0.02 | $< 10^{-10}$ | 0% |
| phi4(FT) | race | Asian & Pacific | -0.03 | 0.08 | $< 10^{-10}$ | 0% |
| phi4(FT) | race | Black | -0.02 | 0.08 | $< 10^{-10}$ | 0% |
| phi4(FT) | race | Hispanic/Latino | -0.02 | 0.08 | $< 10^{-10}$ | 0% |
| phi4(FT) | race | Other/Unknown | 0.19 | 0.10 | $< 10^{-10}$ | 0% |
| phi4(FT) | race | White | -0.03 | 0.08 | $< 10^{-10}$ | 0% |
| phi4(ZS) | age | young adults | -4.28 | 2.88 | $< 10^{-10}$ | 67% |
| phi4(ZS) | age | middle aged adults | -2.28 | 2.51 | $< 10^{-10}$ | 49% |
| phi4(ZS) | age | older adults | 0.15 | 2.54 | $< 10^{-10}$ | 37% |
| phi4(ZS) | age | elderly | 3.36 | 2.82 | $< 10^{-10}$ | 48% |
| phi4(ZS) | gender | F | -0.12 | 0.20 | $< 10^{-10}$ | 2% |
| phi4(ZS) | gender | M | 0.15 | 0.23 | $< 10^{-10}$ | 2% |
| phi4(ZS) | race | Asian & Pacific | -0.02 | 0.30 | $< 10^{-10}$ | 4% |
| phi4(ZS) | race | Black | 0.39 | 0.32 | $< 10^{-10}$ | 6% |
| phi4(ZS) | race | Hispanic/Latino | 0.23 | 0.35 | $< 10^{-10}$ | 6% |
| phi4(ZS) | race | Other/Unknown | -0.03 | 0.30 | $< 10^{-10}$ | 4% |
| phi4(ZS) | race | White | -0.02 | 0.25 | $< 10^{-10}$ | 2% |

Table 11: Detailed demographic sensitivity results, including mean $\Delta\mathbb{E}_{\text{LOS}}$, standard deviation, p-values, and class flip percentages for each model, variable, and demographic class.

| Model | Variable | Class | Mean ΔE | Std ΔE | p-value | % Class Flip |
|---|---|---|---|---|---|---|
| obllm(FT) | age | middle aged adults | -0.03 | 0.25 | $<10^{-10}$ | 0% |
| obllm(FT) | age | older adults | 0.12 | 0.24 | $<10^{-10}$ | 0% |
| obllm(FT) | age | elderly | -0.03 | 0.31 | $<10^{-10}$ | 0% |
| obllm(FT) | gender | F | -0.03 | 0.03 | $<10^{-10}$ | 0% |
| obllm(FT) | gender | M | 0.03 | 0.03 | $<10^{-10}$ | 0% |
| obllm(FT) | race | Asian & Pacific | -0.02 | 0.15 | $<10^{-10}$ | 0% |
| obllm(FT) | race | Black | -0.04 | 0.15 | $<10^{-10}$ | 0% |
| obllm(FT) | race | Hispanic/Latino | -0.07 | 0.15 | $<10^{-10}$ | 0% |
| obllm(FT) | race | Other/Unknown | 0.32 | 0.15 | $<10^{-10}$ | 0% |
| obllm(FT) | race | White | -0.06 | 0.15 | $<10^{-10}$ | 0% |
| obllm(ZS) | age | young adults | -1.39 | 0.90 | $<10^{-10}$ | 15% |
| obllm(ZS) | age | middle aged adults | -0.76 | 0.88 | $<10^{-10}$ | 8% |
| obllm(ZS) | age | older adults | -0.05 | 0.87 | $<10^{-10}$ | 8% |
| obllm(ZS) | age | elderly | 1.56 | 1.16 | $<10^{-10}$ | 35% |
| obllm(ZS) | gender | F | 0.00 | 0.11 | 0.0967 | 1% |
| obllm(ZS) | gender | M | 0.01 | 0.11 | 0.0967 | 2% |
| obllm(ZS) | race | Asian & Pacific | -0.47 | 0.25 | $<10^{-10}$ | 6% |
| obllm(ZS) | race | Black | 0.53 | 0.25 | $<10^{-10}$ | 6% |
| obllm(ZS) | race | Hispanic/Latino | -0.28 | 0.28 | $<10^{-10}$ | 4% |
| obllm(ZS) | race | Other/Unknown | 0.18 | 0.24 | $<10^{-10}$ | 3% |
| obllm(ZS) | race | White | -0.11 | 0.23 | $<10^{-10}$ | 2% |
| meditron(ZS) | age | young adults | -2.63 | 1.27 | $<10^{-10}$ | 66% |
| meditron(ZS) | age | middle aged adults | -1.07 | 1.22 | $<10^{-10}$ | 41% |
| meditron(ZS) | age | older adults | 0.26 | 1.22 | $<10^{-10}$ | 25% |
| meditron(ZS) | age | elderly | 0.93 | 1.20 | $<10^{-10}$ | 37% |
| meditron(ZS) | gender | F | -0.06 | 0.07 | $<10^{-10}$ | 3% |
| meditron(ZS) | gender | M | 0.06 | 0.06 | $<10^{-10}$ | 2% |
| meditron(ZS) | race | Asian & Pacific | 0.03 | 0.14 | $<10^{-10}$ | 4% |
| meditron(ZS) | race | Black | 0.15 | 0.11 | $<10^{-10}$ | 4% |
| meditron(ZS) | race | Hispanic/Latino | 0.25 | 0.13 | $<10^{-10}$ | 9% |
| meditron(ZS) | race | Other/Unknown | -0.01 | 0.12 | $<10^{-10}$ | 3% |
| meditron(ZS) | race | White | 0.03 | 0.10 | $<10^{-10}$ | 3% |
| meditron(ZS) | age | young adults | -2.63 | 1.27 | $<10^{-10}$ | 66% |
| meditron(ZS) | age | middle aged adults | -1.07 | 1.22 | $<10^{-10}$ | 41% |
| meditron(ZS) | age | older adults | 0.26 | 1.22 | $<10^{-10}$ | 25% |
| meditron(ZS) | age | elderly | 0.93 | 1.20 | $<10^{-10}$ | 37% |
| meditron(ZS) | gender | F | -0.06 | 0.07 | $<10^{-10}$ | 3% |
| meditron(ZS) | gender | M | 0.06 | 0.06 | $<10^{-10}$ | 2% |
| meditron(ZS) | race | Asian & Pacific | 0.03 | 0.14 | $<10^{-10}$ | 4% |
| meditron(ZS) | race | Black | 0.15 | 0.11 | $<10^{-10}$ | 4% |
| meditron(ZS) | race | Hispanic/Latino | 0.25 | 0.13 | $<10^{-10}$ | 9% |
| meditron(ZS) | race | Other/Unknown | -0.01 | 0.12 | $<10^{-10}$ | 3% |
| meditron(ZS) | race | White | 0.03 | 0.10 | $<10^{-10}$ | 3% |

Table 12: Extended demographic analysis for additional models. Metrics follow the same format as Table 11.