DeVisE: Behavioral Testing of Medical Large Language Models

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly used in clinical decision support, yet current evaluation methods often fail to distinguish genuine medical reasoning from superficial patterns. We introduce DeVisE (**De**mographics and **Vi**tal **s**igns **E**valuation), a behavioral testing framework for probing fine-grained clinical understanding. We construct a dataset of ICU discharge notes from MIMIC-IV, generating both raw (real-world) and template-based (synthetic) versions with controlled single-variable counterfactuals targeting demographic (age, gender, ethnicity) and vital sign attributes. On DeVisE, we evaluate five LLMs spanning general-purpose and medically fine-tuned variants, under both zero-shot and fine-tuned settings. We assess model behavior via (1) input-level sensitivity–how counterfactuals alter the likelihood of a note; and (2) downstream reasoning–how they affect predicted hospital length-of-stay. Our results show that zero-shot models exhibit more coherent counterfactual reasoning patterns, while fine-tuned models tend to be more stable yet less responsive to clinically meaningful changes. Notably, demographic factors subtly but consistently influence outputs, emphasizing the importance of fairness-aware evaluation. This work highlights the utility of behavioral testing in exposing the reasoning strategies of clinical LLMs and informing the design of safer, more transparent medical AI systems.[1]

1 Introduction

Large language models (LLMs) are increasingly applied to the medical domain, showing strong performance in clinical tasks when fine-tuned on domain-specific data (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-grained metrics such as AUROC and F1 scores, which provide limited insight into whether models perform deep medical reasoning (Van Aken et al., 2021a; MacPhail et al., 2024; Jullien et al., 2024) or rely on shortcuts and spurious correlations. In fact, models that perform well on these benchmarks can still struggle in scenarios requiring fine-grained clinical reasoning (Aguiar et al., 2024; Ceballos-Arroyo et al., 2024), particularly in tasks involving temporal understanding of events in clinical notes (MacPhail et al., 2024; Kougia et al., 2024).

To address this gap, behavioral testing offers a complementary evaluation paradigm. Originating in software engineering, it involves assessing a system's behavior through its input–output responses, without requiring access to its internal workings (Beizer and Wiley, 1996). In Natural Language Processing (NLP), Ribeiro et al. (2020) adapted this approach to evaluate linguistic capabilities across a range of tasks.

In this work, we introduce **DeVisE**, a behavioral testing benchmark based on MIMIC-IV discharge summaries (Johnson et al., 2023) focusing on *minimally differing counterfactuals* in key clinical variables: *demographics* (age, gender, ethnicity) and *vital signs* (heart rate, respiration rate, oxygen saturation, and blood pressure). Our benchmark includes $1,000$ high-quality, manually clinical notes. Behavioral testing in this context enables transparent evaluation of LLMs, helping clinicians and developers assess whether model outputs are grounded in clinically meaningful reasoning, an essential step given that clinical outcomes can depend on subtle variations in patient data such as changes in vital signs (Alghatani et al., 2021; Downey et al., 2017; Herasevich et al., 2022).

Because raw clinical notes are often noisy, filled with abbreviations and domain-specific jargon, we additionally construct synthetic, template-based

[1] We will publicly release all code needed to reproduce our results and the benchmark data.

notes that are noise-free and contain only the variables of interest and their counterfactuals. By comparing model behavior on both raw and template-based clinical notes, we address a relatively unexplored aspect of medical LLM evaluation in behavioral testing setting. To our knowledge, most prior approaches have focused solely on structured templates or synthetic inputs (Van Aken et al., 2021a; MacPhail et al., 2024; Aguiar et al., 2024; Rajagopal et al., 2021).

Our contributions in this study are:

- We introduce **DeVisE** (**De**mographics and **Vi**tal **s**igns **E**valuation), a novel clinical NLP benchmark based on behavioural testing with *minimally differing counterfactuals* across demographics and vital signs clinical variables.
- We create raw and template-based clinical notes and their counterfactuals, and compare how models fare when applied to the raw notes and to template-based clinical notes.
- We compare state-of-the-art LLMs across different dimensions: fine-tuned vs. zero-shot, medical vs. general purpose, reasoning vs. non-reasoning, and given only inputs vs. downstream (see Fig. 1 for more details).

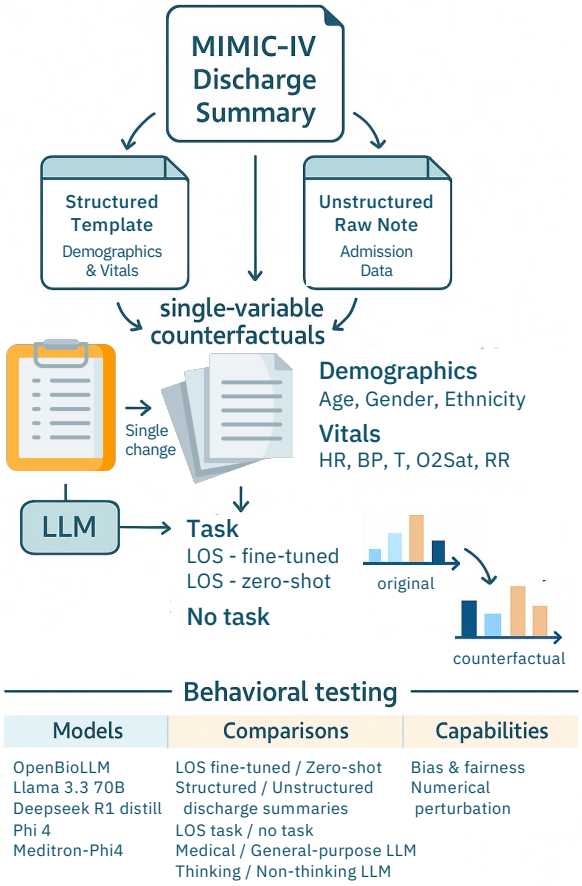


Figure 1: **Overview of DeVisE**. We construct a dataset of 1,000 MIMIC-IV discharge summaries with manually validated single-variable counterfactuals targeting key clinical attributes, including **demographics** (age, gender, ethnicity) and **vital signs** (heart, respiration rate, oxygen saturation, temperature, blood pressure). We compare LLM predictions using both *raw* (noisy, real-world) and *template-based* (clean, synthetic) clinical notes. Behavioral evaluations are conducted across five LLMs, probing their 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?").

2 Related Work

Several studies have explored the limitations of traditional evaluation methods for LLMs in the medical domain, and have proposed frameworks to better understand model behavior. Lee et al. (Lee et al., 2025), analyzed LLM robustness to distribution shifts and missing data in hospital records for triage, finding improved performance over traditional models but evidence of demographic biases, particularly along gender and racial lines.

In the context of demographic variables, van Aken et al. (Van Aken et al., 2021a) showed that models with similar AUROC scores can behave differently when evaluated on finer-grained capabilities such as bias sensitivity through age, gender, and ethnicity, and Zack et al.(Zack et al., 2024) and Zhao et al. (Zhao et al., 2024) further found that GPT-4 and other LLMs often favor majority groups, producing less accurate predictions for minorities and amplifying existing disparities.

Researchers have also studied robustness in clinical tasks by examining whether models produce consistent outputs in response to small input variations. MacPhail et al.(MacPhail et al., 2024) introduced template-based tests for adverse drug event classification, showing inconsistent performance on capabilities like temporal reasoning and negation, even among models with similar performances. Kougia et al.(Kougia et al., 2024) similarly found that biomedical LLMs often fail at event sequencing, affecting reliability in clinical decision-making. Natural Language Inference (NLI) has been used to assess reasoning consistency. Altinok et al. (Altinok, 2024) and Aguiar et al. (Aguiar et al., 2024) introduced contrast sets and system-

atic perturbations to evaluate modelfaithfulness and consistency, revealing substantial variance even among closely related models.

While these studies highlight the limitations of traditional evaluation methods, most focus on structured templates. Behavioral analysis over raw clinical notes, especially with controlled perturbations, remains largely unexplored. To address this gap, we introduce **DeVisE**, a systematic behavioral evaluation framework designed to assess model sensitivity to demographic and vital sign variations in both raw and template-based clinical notes.

3 Methods

DeVisE is a behavioral testing framework designed to evaluate two key capabilities in medical LLMs: sensitivity to demographic variables and to vital signs. Sensitivity to demographics is crucial for mitigating biases, while sensitivity to vitals requires basic numerical reasoning. To probe these capabilities, we construct counterfactual discharge summaries in which only one clinical variable is modified at a time.

3.1 Clinical variables

Demographics. We focus on *gender*, *age*, and *ethnicity*, selected for their availability in clinical notes and documented impact on model bias (Van Aken et al., 2021a; Celi et al., 2022). For example, fairness can be assessed by comparing predicted probabilities across demographic groups (Zhao et al., 2024).

Vital signs. We select *heart rate*, *respiration rate*, *oxygen saturation*, *blood pressure*, and *temperature* based on their clinical importance, availability, and sensitivity in decision making (Alghatani et al., 2021; Downey et al., 2017; Herasevich et al., 2022). We expect small numerical changes in vitals to lead to proportionally small changes in model predictions; large prediction shifts from minor input changes indicate instability and poor robustness.

3.2 Data

Dataset. To create DeVisE, we use MIMIC-IV discharge summaries (Johnson et al., 2023), which document the full course of a patient's hospital stay.

Raw clinical notes. Following Röhr et al. (2024) and Van Aken et al. (2021b), *we extract only patient information available at admission* to avoid data leakage from events that take place after the admission. 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 construct synthetic versions of the notes that include only demographics and vitals. These noise-free templates allow us to evaluate model behavior in a controlled setting (see Figure A.2).

Population statistics. Our test set includes 1,000 admissions, broadly representative of the adult intensive care 201 unit (ICU) 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%). For each variable, we replace non-missing values and generate counterfactuals: 33,005 (blood pressure), 28,650 (heart rate), 22,110 (temperature), 19,250 (respiration rate), and 13,286 (oxygen saturation).

3.2.1 Counterfactuals

We create counterfactuals by systematically modifying one variable per summary, either a demographic attribute or a vital sign, while keeping all other content unchanged. For each variable, values are grouped into clinically meaningful classes based on official guidelines (American Heart Association, 2024; Royal College of Physicians, 2017; World Health Organization, 2016; Society of Critical Care Medicine, 2021, 2015; Infectious Diseases Society of America, 2003) (see Tables 1 and 2). We sample five random values per class from validated ranges and apply these replacements to generate counterfactual notes.

Vital signs mentioned in raw notes often do not align with structured data. Therefore, we extract vital values directly from the *physical exam* section using few-shot prompting with LLaMA 3 70B (Grattafiori et al., 2024). Rule-based approaches were insufficient due to high variability in phrasing; LLM extraction yielded more reliable results, confirmed via manual inspection.

Counterfactuals evaluation. We validate counterfactual consistency using a combination of automated and manual checks. First, we use the GNU diff tool[2] to identify changes between original and counterfactual summaries. Then, we confirm whether the modified value matches the in-

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

Demographic	Classes
Age	18–35: young adults 36–55: middle aged adults 56–75: older adults $\geq$ 76: elderly
Gender	Female Male
Ethnicity	Asian & Pacific Black Hispanic/Latino Other/Unknown White

Table 1: Categories used for demographic counterfactual variables.

Vital Sign	Classes	Range
Heart Rate (bpm)	Very low	$\leq$40
	Low	41–50
	Normal	51–90
	High	91–110
	Very high	111–130
	LTH	$\geq$131
Blood Pressure (mmHg)	Very low	$\leq$ 70/40
	Low	71–89 / 41–59
	Normal	90–119 / 60–79
	Elevated	120–129 / 60–79
	High	130–139 / 80–89
	Very high	140–179 / 90–119
	LTH	$\geq$ 180/120
Respiration Rate (bpm)	Very low	$\leq$ 8
	Low	9–11
	Normal	12–20
	High	21–24
	Very high	$\geq$ 25
Oxygen Saturation (% SpO_2)	LTL	$\leq$ 91
	Very low	92–93
	Low	94–95
	Normal	$\geq$96
Temperature (°F)	LTL	$\leq$ 82.4
	Very low	82.5–89.4
	Low	89.5–94.9
	Normal	95.0–100.2
	High	100.3–103.9
	LTH	$\geq$ 104.0

Table 2: Categorization of vital signs into clinically meaningful ranges for counterfactual generation. **LTL**: life-threateningly low. **LTH**: life-threateningly high.

tended counterfactual. Any unexpected edits are flagged for manual correction. This iterative process improved reliability.

Manual review revealed a 5% error rate in the automatically extracted vitals. Since demographic values were taken from structured data and modified via regular expressions, their error rate was zero. The final test set consists of 1,000 manually verified notes and their associated counterfactuals.

Class	Length of Stay (LOS)
1	$\leq$ 3 days
2	$>$ 3 and $\leq$ 7 days
3	$>$ 7 and $\leq$ 14 days
4	$>$ 14 days

Table 3: Length-of-stay (LOS) class definitions used for downstream prediction tasks.

4 Experimental Setup

We evaluate LLM robustness to minimal edits in demographics and vital signs using two approaches: 1) Measuring changes in log-probabilities between original and counterfactual notes; and 2) Predicting length-of-stay (LOS) from both versions and analyzing shifts in their predicted distributions. We choose LOS as the downstream task due to its clinical relevance for hospital resource planning and patient management (McMullan et al., 2004) and its dependence on multiple patient factors (Zebin et al., 2019; Naemi et al., 2021). Since admission-time vitals are known to influence LOS prediction, this task offers a meaningful testbed for reasoning. We follow prior work (Röhr et al., 2024) to define four LOS classes (see Table 3).

Admission note probabilities. We compute the average log-likelihood per token and analyze the mean and standard deviation of differences between original and counterfactual notes. This reveals model sensitivity and "surprise" to input changes.

Zero-shot and fine-tuned LOS. Models are evaluated both zero-shot and after fine-tuning for 4-way LOS classification. Fine-tuning uses MIMIC-IV admission summaries. We compare general-purpose and medical-domain models to assess whether task alignment improves robustness. Details are in Appendix A.3.

4.1 Evaluation Metrics

We analyze model behavior by comparing predicted LOS distributions for original vs. counter-

factual notes. For each pair, we compute:

- **Jensen–Shannon divergence (JSD)** between the predicted distributions.
- **Expected LOS shift** $\Delta\mathbb{E}_{\text{LOS}}$, calculated as the difference in expected LOS (i.e., probability-weighted average of class durations).

These metrics 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.

Severity scale. We define a severity scale from -4 to $+4$ based on how much a variable's class changes from original to counterfactual (e.g., from "normal" to "very high" = +2). This enables tracking whether model outputs shift in clinically expected directions.

We use the following behavioral metrics:

- **Percentage of correct direction (%Corr Dir).** This metric measures how often the model's $\Delta\mathbb{E}_{\text{LOS}}$ aligns with the direction of change in severity. Specifically, whether an increase in severity led to an increase in predicted LOS and vice-versa.
- **Percentage of monotonicity (%Mono).** This metric tracks whether the predicted $\Delta\mathbb{E}_{\text{LOS}}$ follow a consistent upward/downward trend across severity levels. For each of the 9 severity levels, 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).** While the previous metrics focus on the expected LOS values, this metric looks at prediction stability in terms of class labels. It measures 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.** To understand whether LLMs favor increased or decreased LOS, we calculate the percentage of counterfactuals that resulted in increased ($\%\Delta E^{+}$) versus decreased ($1 - \%\Delta E^{+}$) expected 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 and domains: OpenBioLLM (Ankit Pal, 2024), Meditron3-Phi4 (OpenMeditron, 2025), Phi 4 (Abdin et al., 2024), LLaMA 3 Instruct (Grattafiori et al., 2024), and DeepSeek R1 Distill (Guo et al., 2025).

We categorize the models along three axes used in our analysis: (1) fine-tuned vs. zero-shot, (2) medical-domain vs. general-purpose, and (3) reasoning-oriented vs. non-reasoning. These groupings are used to assess differences in robustness, sensitivity, and behavioral consistency across model types. See Table 7 for model specifications and categorizations.

5 Results

5.1 Vital signs

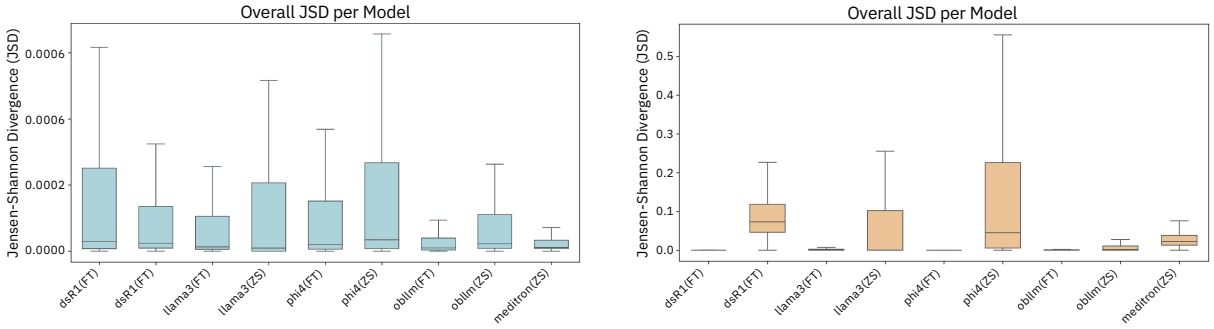
Figure 2 compares the Jensen-Shannon divergence (JSD) across models when perturbed with vital sign counterfactuals. On raw notes, vital changes cause minimal distributional shift (Figure 2a), while template-based notes lead to notably higher shifts, especially in the zero-shot (ZS) setting (Figure 2b).

Fine-tuned (FT) models achieve modest but higher F1 scores than their ZS counterparts (Table 4). However, ZS models show greater behavioral sensitivity, with elevated $\Delta\mathbb{E}_{\text{LOS}}$, %Flip, and %Mono. Flip rates in FT models remain below 1

The relationship between perturbation severity and $\Delta\mathbb{E}_{\text{LOS}}$ (Figure 3) is more clinically coherent for raw notes. While ZS models predict larger changes on templates, they often fail to decrease LOS with decreasing severity, indicating less reliable monotonic behavior.

All models tend to increase LOS predictions under perturbations ($\%\Delta E^{+} > 50$), with the strongest effect in ZS template-based settings (e.g., Meditron: 54% $\rightarrow$ 96%). This pattern indicates models' general inclination to associate changes in vitals with worsening outcomes.

Behavioral differences also emerge without an explicit prediction task. Changes in token-level log-



(a) Overall JSD shift per model (raw notes)

(b) Overall JSD shift per model (template-based notes)

Figure 2: Comparison of JSD per model using raw vs structured notes. Templates-based experiments in ZS setting show a greater JSD than FT and all experiments on raw notes. (For clarity, outliers are hidden).

Model	Setting	Acc	F1	% ΔE^+	Avg ΔE^+ / ΔE^-	Std ΔE^+	Std ΔE^-	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	63	respiration rate
llama3(FT)	raw	0.722	**0.311**	63	0.130 / -0.057	0.356	0.098	0.001	0.006	42	0	63	blood pressure
phi4(FT)	raw	**0.723**	0.273	64	0.126 / -0.096	0.308	0.199	0.001	0.005	41	0	63	blood pressure
obllm(FT)	raw	**0.723**	0.240	63	0.072 / -0.068	0.267	0.179	0.000	0.005	43	0	63	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	88	temperature
llama3(ZS)	raw	0.116	0.119	64	0.138 / -0.146	0.517	0.600	0.006	0.047	42	2	88	blood pressure
phi4(ZS)	raw	0.080	0.082	63	0.146 / -0.156	0.294	0.521	0.002	0.016	42	2	**100**	blood pressure
obllm(ZS)	raw	0.119	0.118	58	0.109 / -0.139	0.200	0.411	0.001	0.006	37	2	75	blood pressure
dsR1(FT)	template	0.722	0.210	63	0.167 / -0.148	0.175	0.185	0.000	0.001	**49**	0	44	respiration rate
llama3(FT)	template	0.722	0.210	66	0.464 / -0.258	0.401	0.267	0.003	0.004	44	0	44	respiration rate
phi4(FT)	template	0.722	0.210	59	0.138 / -0.121	0.216	0.200	0.001	0.002	44	0	44	respiration rate
obllm(FT)	template	0.722	0.210	67	0.271 / -0.194	0.227	0.201	0.001	0.002	47	0	67	oxygen saturation
meditron(ZS)	template	0.327	0.184	**96**	2.198 / -0.429	1.040	0.525	0.029	0.026	35	**73**	89	temperature
dsR1(ZS)	template	0.217	0.113	49	1.091 / **-1.185**	1.318	0.856	0.087	0.059	37	15	78	respiration rate
llama3(ZS)	template	0.258	0.149	76	1.925 / -0.398	3.805	1.334	0.151	0.311	42	18	56	oxygen saturation
phi4(ZS)	template	0.187	0.165	76	**3.679** / -0.985	3.534	1.452	**0.158**	0.228	40	41	89	oxygen saturation
obllm(ZS)	template	0.179	0.098	75	1.033 / -0.311	1.189	0.481	0.009	0.020	31	15	78	oxygen saturation

Table 4: Detailed performance and behavioral summary for each model variant across note types. FT = fine-tuned, ZS = zero-shot. Orange = medical LLM, gray = reasoning LLM, white = general purpose. Highest values are bold, and lowest are underlined.

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

Table 5: Mean Δ log-likelihood per token ($\pm$ std) across models.

likelihoods remain small but consistently scale with severity (Table 5, Figure 4), suggesting models are linguistically surprised by the counterfactuals even when untrained on LOS prediction.

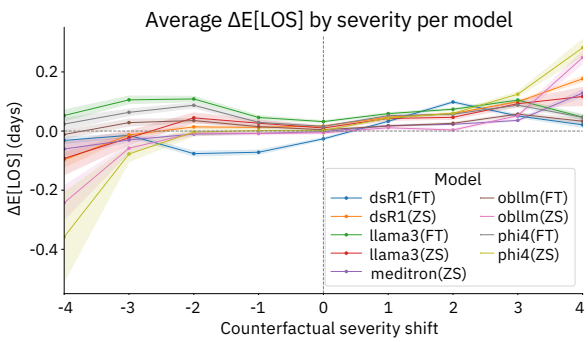
Medical LLMs in the FT setting are more conservative, showing lower JSD and $\Delta\mathbb{E}_{\text{LOS}}$, yet comparable accuracy and directional correctness to general-purpose models. In contrast, medical ZS models outperform on accuracy and F1, albeit with higher flip rates. DeepseekR1 stands out with the highest % correct direction across most settings and is particularly stable in ZS raw notes.

Figure 5 breaks down JSD by vital sign. In raw notes, blood pressure dominates model sensitivity, while in templates, respiration rate is most influential. Some models, like DeepseekR1(FT), maintain consistent vital focus, while others shift across FT/ZS settings.

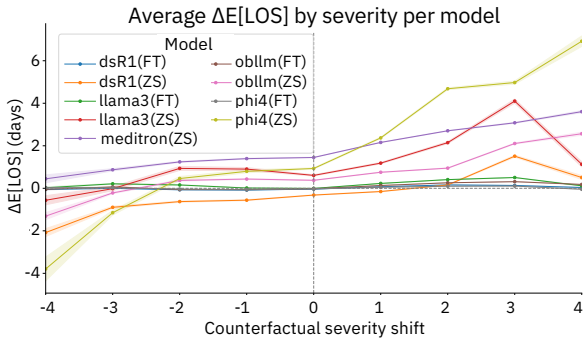
5.2 Demographic variables

Demographic perturbations reveal consistent and often statistically significant effects in the ZS setting with template-based inputs (Figure 6). Among demographic factors, age consistently shows the strongest influence: older age groups are associated with longer predicted LOS across all models. These effects persist despite minimal textual changes, emphasizing the models' sensitivity to demographic shifts.

In contrast, gender and race show smaller average effects, but still impact predictions near deci-



(a) Avg $\Delta\mathbb{E}_{\text{LOS}}$ by severity per model (raw notes)



(b) Avg $\Delta\mathbb{E}_{\text{LOS}}$ by severity per model (template-based notes)

Figure 3: Effect of note modification on $\Delta\mathbb{E}_{\text{LOS}}$ patterns by severity across models using raw vs template-based notes. Raw notes show a more clinically reasonable pattern. Template-based notes show greater magnitude of effect.

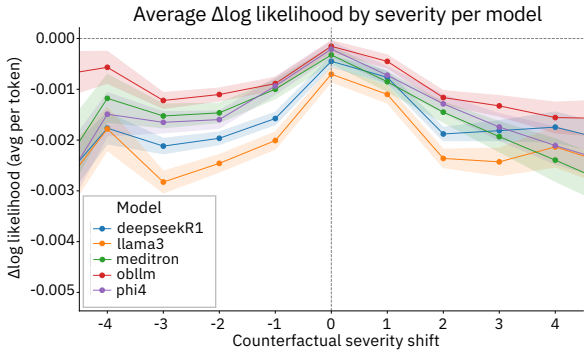


Figure 4: ΔAvg log-likelihood per model (raw notes). There is an increased change in log-likelihood with increasingly positive or negative severity.

sion boundaries, as evidenced by flip rates reaching up to 6% in ZS models. While most model comparisons for gender and race are significant, exceptions include gender in obllm(ZS) and the "Other/Unknown" race category in llama3(ZS).

Model-wise, dsR1(ZS) exhibits the most stable behavior, with the lowest $\Delta\mathbb{E}_{\text{LOS}}$ and minimal flips. Conversely, obllm(ZS) and phi4(ZS) display larger shifts for race and gender, respectively. Notably, race effects tend to be larger than gender effects, and all models predict shorter LOS for female and Black patients compared to other groups. These disparities mirror known societal biases and raise important concerns about fairness in medical LLMs.

In the FT setting, models are much less responsive to demographic changes: $\Delta\mathbb{E}_{\text{LOS}}$ is near zero, and no flips are observed (Figure 8). This suggests that task-specific fine-tuning may inadvertently suppress meaningful demographic sensitivity ,whether beneficial or harmful, and highlights the need for fairness-aware training objectives.

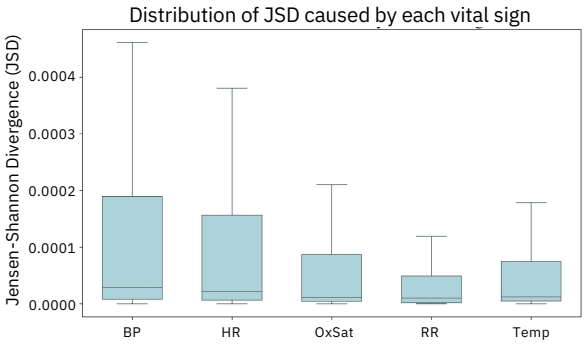
For comprehensive results disaggregated by model, demographic variable, and class, see Tables 10 and 11.

6 Discussion

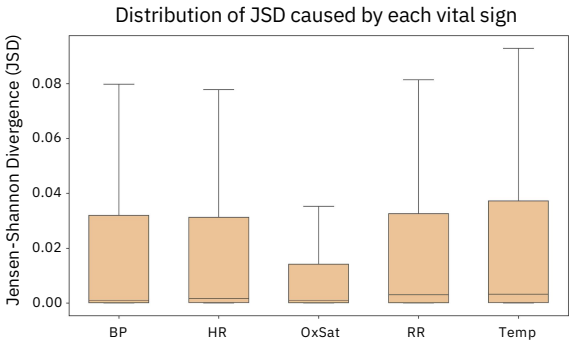
Fine-tuned vs Zero-shot. Despite lower accuracy, zero-shot models exhibit clearer behavioral trends in both JSD and LOS shifts across severity levels (Table 4). This suggests that, without task-specific tuning, they may rely more on generalizable reasoning patterns, e.g., associating low oxygen saturation 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 always improves robustness, raising questions about when it instead reinforces heuristic behaviors.

Medical vs General-purpose. Domain-specific models, both fine-tuned and zero-shot, reacted less strongly to counterfactuals than general-purpose ones, as indicated by lower JSD and $\Delta\mathbb{E}_{\text{LOS}}$ value(Tables 8, 9), possibly due to more conservative priors learned from biomedical corpora. However, they achieved higher %Mono, suggesting closer alignment with clinical expectations. Interestingly, zero-shot medical models outperformed general ones in accuracy despite showing more prediction flips, likely reflecting better domain priors.

Reasoning vs Non-reasoning. DeepSeekR1 demonstrated strong task-aligned and clinically coherent behavior, particularly in its fine-tuned variant. It achieved high %Corr Dir with minimal prediction flips, especially on raw notes (Figure 3), positioning it between the conservative medical models and the more volatile general-purpose ones.



(a) JSD shift per vital (raw notes).



(b) JSD shift per vital (template-based notes)

Figure 5: Comparison of JSD per vital sign using raw vs template-based notes across all models. (For clarity, outliers are hidden). Low magnitude of effects with a subtle different ranking between raw and template-based notes.

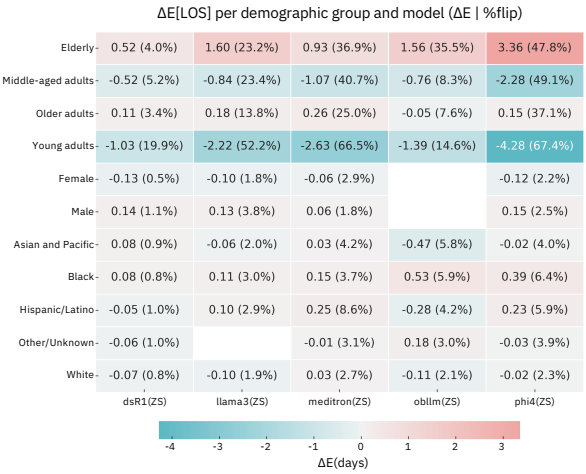


Figure 6: Heatmap $\Delta\mathbb{E}_{\text{LOS}}$ by demographic classes per model (zero-shot) All comparisons between variables are significant except the tiles left blank. Age variable shows the biggest effects and flips.

These findings suggest that reasoning-focused pretraining can offer a balance of robustness and sensitivity, making it a promising foundation for developing clinically reliable LLMs.

Task vs No-task. Even in the absence of a downstream task, models responded systematically to increasing counterfactual severity. Log-likelihood shifts follow a pyramid-like pattern across severity levels (Figure 4), reflecting that models perceive the perturbations as increasingly surprising at the language modeling level. This supports the use of task-agnostic behavioral probes as an early diagnostic signal.

Template-based vs Raw Notes. Template-based notes resulted in noisier and less coherent LOS predictions (Fig.3). 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 Tables 4, 5) highlights ongoing uncertainty in model behavior.

Demographic Variables. Demographic attributes influenced model predictions. While average LOS shifts were small for gender and race, frequent class flips near decision boundaries (Fig. 6) suggest potential fairness concerns. Consistent underprediction of LOS for women and black patients across all models highlights the importance of including demographic evaluation in clinical LLM benchmarks. These patterns likely reflect biases encoded during pretraining, even in zero-shot models.

7 Conclusion

These results highlight the importance of behavioral evaluation in understanding how medical LLMs respond to perturbations. We find that zero-shot models can display surprisingly coherent reasoning patterns, while fine-tuned models tend to favor conservative and stable outputs, sometimes at the expense of sensitivity to clinically meaningful changes. Demographic variables subtly but consistently influence predictions, underscoring the need for fairness-aware evaluations. Future research should explore targeted fine-tuning strategies that preserve reasoning ability while improving calibration and fairness, as well as the development of diagnostic tools to detect and mitigate model bias in real-world clinical settings.

Limitations

This work does not cover an extensive range of model capabilities, other relevant aspects such as temporal reasoning are not included. Additionally, the fine-tuning data for length-of-stay (LOS) prediction was imbalanced, which may have affected model performance. Finally, 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.

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.

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*.

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*.

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.

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. 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.

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.

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 Full Population Statistics

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

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

A.5 Results

A.5.1 Model Specifications

Table 7 summarizes the LLMs used, including model type, domain, and background info on training.

A.5.2 Model Group Comparisons

Summary statistics and pairwise significance tests across model groups and settings (FT vs. ZS; template vs. raw) are available in Tables 8 and 9.

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

Model	Setting	Acc	F1	%ΔE^+	Avg ΔE^+ / ΔE^-	Std ΔE^+	Std ΔE^-	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 8: 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..

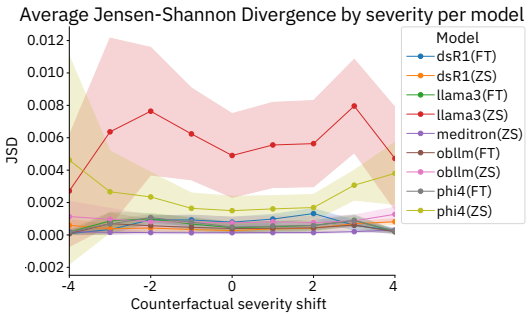


Figure 7: Jensen-Shannon divergence (JSD) across counterfactual severity levels for each model on raw clinical notes. Higher severity generally results in larger distributional shifts.

A.5.3 JSD by Severity (Raw Notes)

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

A.5.4 Demographic variables

Full demographic bias analysis is presented in Tables 10 and 11. 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.

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 9: 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.

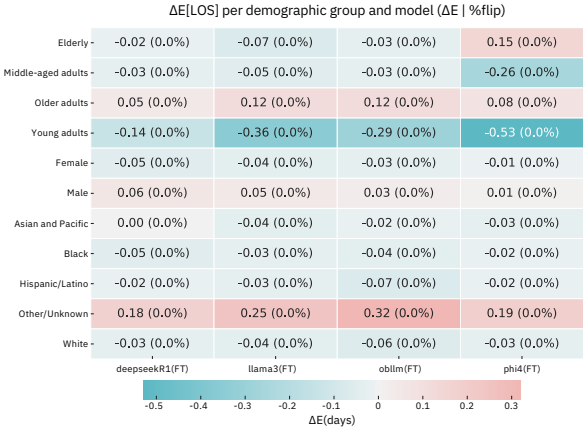


Figure 8: Mean change in expected length-of-stay ($\Delta\mathbb{E}_{\text{LOS}}$) across demographic subgroups for fine-tuned models. All models consistently predict shorter LOS for females and Black patients.

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 10: 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 11: Extended demographic analysis for additional models. Metrics follow the same format as Table 10.