How Robust Are Large Language Models for Clinical Numeracy? An Empirical Study on Numerical Reasoning Abilities in Clinical Contexts

Minh-Vuong Nguyen[◇] **Fatemeh Shiri**[◇] **Zhuang Li**[◇] **Karin Verspoor**[◇]
[◇] School of Computing Technologies, RMIT University, Australia
{vuong.nguyen2, first.last}@rmit.edu.au

Abstract

Large Language Models (LLMs) are increasingly being explored for clinical question answering and decision support, yet safe deployment critically requires reliable handling of patient measurements in heterogeneous clinical notes. Existing evaluations of LLMs for *clinical numerical reasoning* provide limited operation-level coverage, restricted primarily to arithmetic computation, and rarely assess the robustness of numerical understanding across clinical note formats. We introduce **ClinicNumRobBench**, a benchmark of 1,624 context–question instances with ground-truth answers that evaluates four main types of clinical numeracy: *value retrieval*, *arithmetic computation*, *relational comparison*, and *aggregation*. To stress-test robustness, ClinicNumRobBench presents longitudinal MIMIC-IV vital-sign records in three semantically equivalent representations, including a real-world note-style variant derived from the Open Patients dataset, and instantiates queries using 42 question templates. Experiments on 17 LLMs show that *value retrieval* is generally strong, with most models exceeding 85% accuracy, while *relational comparison* and *aggregation* remain challenging, with some models scoring below 15%. Fine-tuning on medical data can reduce numeracy relative to base models by over 30%, and performance drops under note-style variation indicate LLM sensitivity to format. ClinicNumRobBench offers a rigorous testbed for clinically reliable numerical reasoning[1].

1 Introduction

Large Language Models (LLMs) have exhibited rapid exploration of medical applications such as clinical question answering and decision support (Bedi et al., 2025). Yet safe clinical use critically requires LLMs to handle patient measurements reliably, which depends on both *clinical numeracy*, executing basic numerical operations over clinical measurements, and *clinical robustness*, preserving numerical correctness under semantically equivalent variations in how the same measurements are documented across clinical notes.

First, despite substantial progress on mathematical reasoning (Shao et al., 2024), prior numeracy evaluations show that LLMs still exhibit persistent weaknesses in fundamental operations, including *value retrieval*, *arithmetic computation*, *relational comparison*, and *aggregation* (Li et al., 2025; Mahendra et al., 2025). Meanwhile, clinical text is numerically rich, with numbers appearing in most documents across widely used clinical datasets (Mahendra et al., 2024), and these same operations are repeatedly required in clinical workflows over patient measurements, making numerical reliability a prerequisite for safe deployment.

Second, clinical numeracy differs from canonical mathematical word problems with systematic symbolic structure. Clinical measurements are embedded in diverse and noisy contexts, and the same value may appear in structured fields, semi-structured templates, or free-text notes with varied surface forms (e.g., "BP: 120/80" vs. "blood pressure of 120 over 80") (Rothman et al., 2008). As a result, even when an LLM succeeds in one presentation, its numerical behavior may be brittle under semantically equivalent documentation variants.

A clinically meaningful evaluation of LLMs must therefore test both fine-grained numerical operation correctness and robustness under representational shifts where the underlying measurements remain unchanged. However, existing datasets do not jointly address these requirements. General numeracy benchmarks (Mahendra et al., 2024, 2025; Li et al., 2025) are mostly non-clinical, while clinical benchmarks often cover only a subset of operations and omit robustness testing. For instance, MedCalc-Bench (Khandekar et al., 2024) is derived from patient notes but mainly evaluates

[1]Code and data URL are available on https://github.com/MinhVuong2000/ClinicNumRobBench

Findings of the Association for Computational Linguistics: ACL 2026, pages 18113–18129

arithmetic computation and Electromyogram Table Mart (ETM) (Long et al., 2025) focuses on table-to-text diagnosis generation rather than operation-level clinical numeracy benchmarking. Neither consider robustness under note-style variation. A recent study (Gourabathina et al., 2025) examined LLMs' sensitivity to non-clinical input perturbation, with altered patient attributes causing inconsistent treatment recommendations and demographic disparities. While the work focuses on the fairness of LLMs with the change of non-clinical input, the robustness of LLMs on diverse clinical documentation formats remains unexplored, especially numerical reasoning being fundamental to clinical decision-making.

To address this gap, we introduce **ClinicNumRobBench**, a benchmark of 1,624 context–question instances for fine-grained clinical numeracy evaluation and robustness testing, together with a scalable synthetic construction pipeline. Grounded in real-world MIMIC-IV vital-sign records and demographics (Johnson et al., 2023a,b, 2024), our pipeline automatically constructs longitudinal patient contexts by sampling and assembling records, then renders the same underlying measurements in three semantically equivalent context representations, from structured formats to realistic natural-language variants. We instantiate 42 question templates to generate queries probing *value retrieval*, *arithmetic computation*, *relational comparison*, and *aggregation* operations, and compute ground-truth answers programmatically. Using ClinicNumRobBench, we evaluate 17 LLMs across three categories and report three findings in task accuracy, medical fine-tuning negative impact, and context robustness.

Our main contributions are:

- We introduce **ClinicNumRobBench**, a benchmark of 1,624 instances for evaluating clinical numeracy across 4 operations and robustness across 3 context representation formats.
- We propose a scalable construction pipeline grounded in MIMIC-IV longitudinal records that automatically builds the benchmark with minimal human effort.
- Experiments on 17 LLMs reveal persistent weaknesses in comparison and aggregation, potential numeracy degradation from medical fine-tuning, and strong sensitivity to clinical note representational shifts.

2 Related Work

Clinical numeracy evaluation. Most numeracy benchmarks evaluate LLMs using general mathematics questions, ranging from grade school arithmetic to Olympiad level problems (Cobbe et al., 2021; Hendrycks et al., 2021; Li et al., 2025). Despite strong performance on difficult problems, LLMs can still fail simple operations such as comparison or multiplication (Li et al., 2025; Mahendra et al., 2025). Clinical text is numerically rich, yet widely used medical benchmarks such as MedQA, MedXpertQA, and MedBullets (Jin et al., 2020; Zuo et al., 2025; Chen et al., 2025a) assess broad clinical knowledge and reasoning, making it hard to isolate numeracy. Mahendra et al. (2024) identified prevalent types of numerical information in clinical documents, highlighting the lack of systematic resources in the clinical domain for numerical understanding and reasoning. Recently targeted studies on medical calculations and table interpretation (Khandekar et al., 2024; Long et al., 2025) focus on limited formulas or modalities and do not systematically test value retrieval, arithmetic computation, relational comparison, and aggregation.

Robustness to documentation variation. LLMs are sensitive to small changes in prompt wording or formatting (Zhao et al., 2021; Zhuo et al., 2024; Arora et al., 2025). This has motivated robustness benchmarks such as GSM Plus (Li et al., 2024), FinBias (Mehrotra et al., 2025), and CARES (Chen et al., 2025b). In the medical domain, robustness studies mainly target knowledge-oriented adversarial prompts (Ness et al., 2024), or non-clinical input perturbations affecting clinical recommendations (Gourabathina et al., 2025), and rarely examine numerical robustness, even though clinical measurements appear in structured EHR fields, semi-structured templates, and free text notes. Reliable deployment therefore requires testing whether numerical accuracy holds under semantically equivalent documentation variants.

3 ClinicNumRobBench

In this section, we present the **Clinic**al **Num**eracy **Rob**ustness **Bench**mark (**ClinicNumRobBench**) and its construction pipeline, which is used to evaluate numerical reasoning of LLMs in clinical settings and their robustness to semantically equivalent documentation variants. Each instance consists of a patient *context*, a *question*, and an

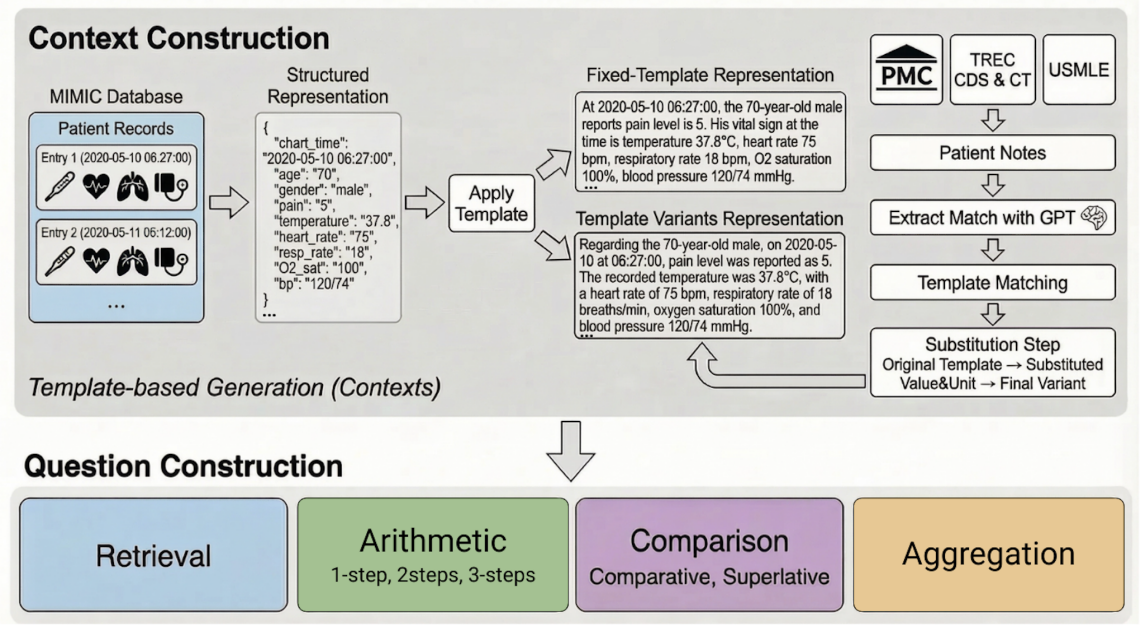

Figure 1: Overview of benchmark construction.

exact ground-truth answer. ClinicNumRobBench integrates (i) multiple clinical context representations and (ii) template-based tasks that target core numerical operations. We then detail context construction, task and question generation, and dataset statistics. A detailed comparison with existing benchmarks, particularly the arithmetic-focused MedCalc-Bench, appears in Table 2 and Appendix B.

3.1 Context Construction

Real world clinical measurements appear in heterogeneous formats, from structured *representations* (e.g., JSON, XML, tables) in EHR systems to unstructured natural language in clinical notes and physician narratives. Because documentation practices vary widely, LLM numeracy behavior may be sensitive to surface form changes. To measure robustness under such representational shifts, we construct three semantically equivalent context *representations*: (i) a structured (JSON like) format, (ii) a fixed natural language template, and (iii) realistic note style variants derived from clinical notes.

We construct contexts from longitudinal vital-sign records and demographics, and render the same underlying measurements into the three semantically equivalent representations described above. Concretely, we (1) extract patient profiles from `hosp/patients` and `icu/icustays` in MIMIC-IV (Johnson et al., 2024), (2) link them to vital-sign records in MIMIC-IV ED (Johnson et al., 2023a), (3) perform stratified sampling over age and gender to obtain 200 patients with 1–50 records each, and (4) instantiate each representation by filling structured and natural-language templates with the resulting values. Each record contains a timestamp, demographics, and clinical measurements, including vital signs and pain level.

3.1.1 Structured Representation

We serialize each record as a JSON-like key–value object and represent a patient context as a chronological list of these objects. Keys are fixed field names, including `chart_time`, demographic fields, and vital-sign attributes shown in Figure 1, and values are the corresponding measurements from the underlying records. This representation mirrors common structured EHR exports and tests whether LLMs can reason over structured clinical data without relying on natural-language cues.

3.1.2 Fixed-template Representation

Prior work has converted structured records into natural language form through verbalization (Nguyen et al., 2024; Oliveira et al., 2025). We adopt the same idea to construct our second context representation. Using a single fixed template enables controlled comparison, since it holds the linguistic form largely constant while varying only the underlying values. This design helps isolate the effect of light linguistic wrapping, such as function words and sentence structure, and quantify robust-

ness to modest representational shifts.

For each patient, we verbalize all records in chronological order. Each record is rendered using the fixed template below:

Record verbalized fixed template

At ⟨charttime⟩, the ⟨age⟩-year-old ⟨gender⟩ reports a pain level of ⟨pain⟩. Vital signs: temperature ⟨temperature⟩°C, heart rate ⟨heartrate⟩ bpm, respiratory rate ⟨resprate⟩ breaths/min, O_2 saturation ⟨o2sat⟩%, blood pressure ⟨sbp⟩/⟨dbp⟩ mmHg.

For patients with multiple records, we concatenate the verbalized records using a newline character \n. For readability, the first record states age and gender explicitly, while subsequent records use pronouns such as "his" or "her". A complete example is shown in Figure 1.

3.1.3 Realistic Variant Templates Representation

Fixed templates cannot capture the linguistic diversity of real clinical documentation.To approximate note-style variability while keeping the underlying measurements unchanged, we derive variant templates from real patient notes and then fill them with the same structured vital-sign values.

We construct these note-style variants as follows:

- **Collect candidate notes.** We source notes from the Open Patients dataset (Khandekar et al., 2024), which contains 180K notes drawn from multiple clinical corpora, including **T**ext **RE**trieval **C**onference (TREC) Clinical Decision Support/Clinical Trials tracks, MedQA-USMLE, and PMC-Patients.
- **Filter for numerically rich notes.** We keep notes that contain at least five numbers and a `vital sign` keyword, yielding 15K candidates.
- **Extract vital-sign mentions.** Using `gpt-4.1-mini`, we extract five vital-sign attributes (temperature, heart rate, respiratory rate, oxygen saturation, and blood pressure). For each attribute, we output a JSON object containing the raw mention span, numeric value, and unit. Missing fields are recorded as `None`.
- **Validate extraction quality.** We require each extracted raw span to appear in the original note, and re-run extraction until this condition is satisfied. This yields 2,300 notes with validated mentions of all five vital signs.
- **Form variant templates.** For each validated note, we locate the extracted mention indices, take the minimal span covering all mentions, and replace each numeric value and unit with placeholders `%v` and `%m`.
- **Fill templates with record values.** We apply the resulting templates to structured vital-sign records by filling placeholders with record values, producing diverse note-style context variants.
- **Assemble patient contexts.** For each patient, we order the resulting note-style records chronologically and concatenate them to form the temporal narrative used as input for all benchmark tasks.

The extraction prompt and template categories with their frequencies are provided in the Appendix (Figures 4 and 6, respectively). An end-to-end pipeline is shown in Figure 1.

3.2 Task Design and Question Generation

ClinicNumRobBench evaluates four clinical numeracy operation types: **value retrieval**, **arithmetic computation**, **relational comparison**, and **aggregation**. We create instances for each operation using template based questions whose answers are computed deterministically from the underlying measurements. Each template specifies (i) the relevant record or records in a patient's longitudinal context, (ii) the required numerical operation, and (iii) an exact ground truth answer. For retrieval, arithmetic computation, and aggregation tasks, answers are formatted as numeric values (integers or decimals rounded to two decimal places). For comparison tasks, answers are formatted as timestamp strings (`YYYY-MM-DD HH:MM`).

To instantiate a template, we first select the relevant record(s) from the structured vital sign sequence according to the template's constraints (for example by timestamp, by position, or by a threshold). We then fill the template slots (such as `<charttime>`, `<vital_parameter>`, or a threshold) with the corresponding values extracted from these records and compute the answer deterministically using the specified operation. Thresholds are selected based on their clinically meaningful cutoffs. The resulting filled question is paired with the same patient context rendered in each of the three context representations (§3.1). Table 1 reports the number of instances and templates per task type. Table 10 in the Appendix provides further details by subtask. Table 8 depicts the alignment of benchmark tasks with clinical numerical information types (Mahendra et al., 2024).

Task	Subtask	#Templates	#Questions
Retrieval	–	6	240
Arithmetic	1 step	4	200
	2 steps	4	200
	3 steps	2	200
Comparison	Comparative	11	191
	Superlative	12	192
Aggregation	–	3	201
Total		**42**	**1,624**

Table 1: Number of templates and questions per task.

3.2.1 Value Retrieval

Value retrieval involves the location and extraction of a single measurement from the patient context given a timestamp. We generate queries over six vital parameters: temperature, heart rate, respiratory rate, oxygen saturation, systolic blood pressure, and diastolic blood pressure. To cover both early and later observations, we sample target records at fixed positions in the timeline (for example the first and fifth records when available) and fill the template with the corresponding measurement and timestamp. The ground truth answer is simply the value of the requested vital sign at that timestamp.

3.2.2 Arithmetic Computation

Arithmetic computation relates to clinically motivated arithmetic over one or more measurements in the context. We group questions into one-step, two-step and three-step subtasks, defined by the number of arithmetic operations involved. For each instance, we select the required record(s) by timestamp or position, substitute their values into the arithmetic expression defined by the template, and compute the exact numeric result as the answer (e.g., $\mathrm{MAP} = \mathrm{DBP} + \frac{1}{3}(\mathrm{SBP} - \mathrm{DBP})$, where MAP, SBP and DBP denote mean arterial pressure, systolic and diastolic blood pressure, respectively). The formulas utilized were derived from the intersection between vital sign information and the 50 most frequently accessed calculators available on MDCalc[2].

3.2.3 Relational Comparison

Relational comparison evaluates ordering and threshold reasoning over longitudinal measurements, including identifying extreme values. We design two query types: (i) *comparative* queries that return the first timestamp at which a vital sign exceeds (`higher`) or falls below (`lower`) a threshold, and (ii) *superlative* queries that return the timestamp of the maximum (`highest`) or minimum (`lowest`) value of a vital parameter. In each case, the template specifies the target vital sign and, for comparative queries, the threshold, and the ground truth answer is the timestamp (YYYY-MM-DD HH:MM) of the matching record.

3.2.4 Aggregation

Aggregation requires combining information across the longitudinal context by counting how many records satisfy a clinical condition over time. Aggregation underlies clinical monitoring patterns such as counting occurrences of abnormal vitals. For each instance, we scan all records, count the number of matches (sometimes after computing a derived quantity like a shock index), and return this integer as the answer.

3.3 Data Comparison and Analysis

Table 1 summarizes ClinicNumRobBench, which comprises 42 templates and 1,624 instances spanning four clinical numeracy operations: retrieval (240), arithmetic (600), comparison (383), and aggregation (201). Arithmetic is evenly distributed across 1-, 2-, and 3-step problems (200 instances each), and comparison is balanced between comparative and superlative queries (191 vs. 192), enabling fine-grained analysis of threshold versus extremum reasoning.

Table 2 contrasts ClinicNumRobBench with MedCalc-Bench, the closest benchmark to ours for clinical numeracy. MedCalc-Bench focuses on arithmetic in a single note-style format, whereas ClinicNumRobBench expands coverage to retrieval, comparison, and aggregation, and provides three semantically equivalent context representations for robustness analysis. The structured and fixed-template representations yield shorter contexts (551–745 context tokens) and lower lexical diversity (Root TTR 4.28–5.51), serving as controlled baselines. In contrast, the note-style variant is substantially longer (1,480 context tokens) and more lexically diverse (Root TTR 12.56–12.87). Compared with MedCalc-Bench (context tokens 613; Root TTR 10.71), our note-style variant also exhibits greater lexical diversity. Since measurements and questions are held fixed across representations, performance differences can be more directly attributed to representational shift.

[2]MDCalc: https://www.mdcalc.com/#Popular

Benchmark	#Inst.	Tasks	Ctx Rep.	Ctx Tok.	Q Tok.	Root TTR
MedCalc-Bench	1,047	arithmetic	note-style	613	41	10.71
ClinicNumRobBench	1,624	retrieval;	structured	745	23	4.93–5.49
(Ours)		arithmetic;	fixed template	551	23	4.28–5.51
		comparison; aggregation	note-style variant	1480	23	12.56–12.87

Table 2: Comparison between ClinicNumRobBench and MedCalc-Bench. Each ClinicNumRobBench row corresponds to one context representation (structured, fixed template, or note-style variant). **Ctx Tok.** and **Q Tok.** denote the average number of context and question tokens, respectively. **Root TTR** denotes the lexical diversity range of the context across all four tasks (root type–token ratio); see Appendix C for details.

4 Experiments

We use ClinicNumRobBench to benchmark LLMs on four core numerical skills and their robustness across three clinical context representations.

Evaluation. We report the accuracy for each task, defined as accuracy $= \frac{\text{match}}{\text{match+unmatch}}$.

For all the output types, the predicted answer should exactly match the ground truth answer. To ensure robust evaluation despite minor formatting variations in LLM numeric outputs, we apply the following post-processing steps: (1) we extract numeric values using the regular expression pattern `-?\d*\.?\d+`, and (2) we normalize trailing zeros by removing .0+ suffixes. This preprocessing handles common LLM output variations, such as appending .0 to integers or including measurement units, while preserving the semantic correctness of predictions. Correctness requires the exact post-processed match with ground truth.

Evaluated Models. We evaluated 17 widely used and efficient LLMs that span three categories: medical-domain models, their corresponding base models, and selected general-purpose LLMs.

- **Medical LLMs:** MedGemma (Sellergren et al., 2025), MediPhi (Corbeil et al., 2025), Meditron3-Llama3.1 (Chen et al., 2024b), Meditron3-Qwen2.5, UltraMedical (Zhang et al., 2024), Huatuo-o1 (Chen et al., 2024a).
- **General Base LLMs:** gemma-3-4b-it (base of MedGemma) (Team et al., 2025), phi-3.5-mini (base of MediPhi) (Abdin et al., 2024), Qwen-2.5-7B (base of Meditron3-Qwen2.5) (Yang et al., 2025), Llama-3.1 (base of Meditron3-Llama3.1, Huatuo-o1, and UltraMedical) (Dubey et al., 2024).
- **General LLMs:** Qwen3-8B, DeepSeek-R1-Distill-8B (DeepSeek-AI, 2025), GPT-4.1-mini, GPT-5 (Achiam et al., 2023), gemma-3-27b-it, and Llama-3.3 70B.

All models are evaluated using the default inference hyperparameters specified in their original releases. For models that provide both standard and reasoning modes (e.g., Qwen3, Phi-3.5), we evaluate each mode accordingly. Following standard practice in numeracy evaluation, we use zero-shot chain-of-thought prompting (figure 5) for all experiments, including those run in reasoning mode.

4.1 Main Results

Table 3 reports the performance of LLMs on the four clinical numeracy tasks across a range of clinical context representations. Overall, models perform strongly on retrieval, but substantial challenges remain in clinical numerical reasoning, particularly for comparison and aggregation.

Finding 1: While retrieval succeeds consistently, comparison and aggregation show critical weaknesses. Retrieval achieves the highest accuracy across nearly all models: most exceed 85%, and some reach 100% (e.g., Gemma, Qwen3, and GPT-4.1-mini). This indicates that LLMs are generally reliable at extracting simple numerical values from clinical contexts. In contrast, arithmetic performance is much more variable, ranging from 18.67% to 97.50%. Notably, MedGemma/Gemma performs strongly relative to both larger models (e.g., Llama-3.1 and Huatuo-o1) and same-sized competitors (e.g., MediPhi/Phi-3.5-mini). However, with the exception of Qwen3 in reasoning mode and GPT-4.1-mini, most models degrade substantially on comparison and aggregation. Comparison exhibits the largest variance among medical models, while several non-medical reasoning models remain weak on aggregation. Most strikingly, some models achieve near-zero accuracy on comparison (e.g., MedGemma 1.56% and Med-

LLMs	Size	Retrieval			Arithmetic			Comparison			Aggregation		
		Struct.	Fixed	Variant	Struct.	Fixed	Variant	Struct.	Fixed	Variant	Struct.	Fixed	Variant
Medical													
medgemma-it	4B	97.50	93.75 -3.75	90.83 -6.17	71.33	73.17 +1.84	68.83 -2.50	1.56	2.35 +0.79	1.83 +0.27	15.92	21.87 +5.95	15.42 -0.50
MediPhi	4B	94.58	93.33 -1.25	86.67 -7.91	22.00	18.67 -3.33	16.50 -5.50	29.50	29.76 -0.26	20.89 -8.61	15.92	26.37 +10.45	13.93 -1.99
Meditron3-Qwen2.5	7B	98.33	98.33 +0.00	90.00 -8.33	83.33	82.67 -0.66	73.50 -9.93	36.29	47.52 +11.23	34.20 -2.09	50.25	51.74 +1.49	44.28 -7.46
Meditron3-Llama	8B	75.83	79.58 +3.75	73.75 -2.08	43.33	53.50 +10.17	46.50 +3.15	2.35	4.18 +1.83	3.13 +0.78	15.42	16.92 +1.50	11.94 -3.48
UltraMedical	8B	89.58	91.25 +1.67	66.25 -23.33	66.67	71.33 +4.66	42.67 -24.00	27.23	33.68 +6.45	18.85 -8.38	31.34	30.85 -0.49	15.42 -15.92
Huatuo-o1	8B	97.08	91.67 -5.41	82.92 -14.16	59.83	57.50 -2.33	52.17 -7.66	49.61	55.35 +5.66	27.94 -21.67	35.32	35.32 +0.00	25.87 -9.45
Base													
gemma-3-it	4B	**100.00**	97.08 -2.92	92.08 -9.92	79.17	81.33 +2.16	73.00 -6.17	63.45	61.88 -1.75	40.47 -22.98	41.79	43.78 +1.99	30.85 -10.94
Phi-3.5-mini-reasoning	4B	98.33	98.75 +0.42	89.17 -9.16	75.83	79.33 +3.50	69.83 -6.00	53.26	57.70 +4.44	46.82 -6.44	55.22	59.70 +4.48	38.36 -16.86
Qwen-2.5	7B	**100.00**	99.17 -0.83	89.58 -10.42	89.17	90.33 +1.16	82.33 -6.84	67.36	72.33 +4.97	52.74 -14.62	60.20	58.21 -1.99	49.25 -10.95
llama-3.1	8B	95.00	96.67 +1.67	88.75 -6.25	60.83	70.50 +9.67	66.33 +5.50	18.36	32.12 +13.67	25.51 +7.15	15.92	29.85 +13.93	20.90 +4.98
General													
DeepSeek-R1-Distill	8B	97.08	89.17 -7.91	85.00 -12.08	63.50	66.67 +3.17	52.33 -11.17	63.97	61.10 -2.87	42.97 -21.00	42.79	34.33 -8.46	30.35 -12.44
Qwen3	8B	99.58	**100.00** +0.42	90.42 -9.16	94.67	91.50 -3.17	82.67 -12.00	71.28	74.93 +3.65	59.61 -11.67	76.62	76.62 +0.00	60.22 -16.40
GPT-4.1-mini	N/A	**100.00**	**100.00** +0.00	96.25 -3.75	96.67	96.33 -0.34	92.50 -4.17	**95.30**	95.04 -0.26	83.03 -12.27	**82.59**	82.09 -0.50	71.64 -10.95
Qwen3-reasoning	8B	**100.00**	**100.00** +0.00	87.92 -12.08	97.17	97.50 +0.33	94.33 -2.84	94.33	**95.30** +0.97	77.28 -17.05	76.62	**82.59** +5.97	63.18 -13.44
gemma-3-it	27B	**100**	**100** +0.00	94.58 -5.42	95.20	96.75 +1.55	92.83 -2.37	78.61	79.51 +0.90	68.26 -10.35	60.15	58.12 -2.03	49.18 -10.97
llama-3.3	70B	99.58	**100** +0.42	96.25 -3.33	96.00	96.33 +0.33	92.17 -3.83	93.74	95.04 +1.30	79.37 -14.37	73.63	75.12 +1.49	61.69 -11.94
GPT-5	N/A	**100**	**100** +0.00	98.75 -1.25	97.33	**98.00** +0.67	93.5 -3.83	86.67	84.75 -1.92	72.22 -14.45	78.61	79.60 +0.99	62.19 -16.42

Table 3: Accuracy (%) of LLMs on four clinical numeracy tasks. Struct., Fixed, and Variant denote the three context representations. +x.xx, -x.xx indicates the accuracy increase and decrease relative to the Struct. score in the same row, respectively. The suffix "-reasoning" indicates that the model is evaluated with its reasoning mode enabled.

itron3 2.35%), and Llama-3.1 falls below 30% on aggregation. These failures suggest that operations requiring global reasoning across multiple values, such as comparison and aggregation, remain a key bottleneck for clinically reliable numeracy. This dramatic failure mode suggests that comparison and aggregation reasoning, particularly operations requiring global reasoning across multiple values, remains a critical weakness in clinical numerical understanding, even for models that excel at retrieval and basic calculations.

Finding 2: Medical fine-tuning erodes reasoning-based LLMs' numerical reasoning. General-purpose LLMs consistently outperform medical-specialized models across all tasks, with the largest gaps on comparison and aggregation. This pattern suggests that medical fine-tuning can reduce numerical reasoning capability. For example, MediPhi attains only 22% on arithmetic, compared to 75.83% for its base model Phi-3.5-mini; it also shows large drops on comparison (-23.76%) and aggregation (-39.3%). We observe similar trends for other medically adapted models, indicating that standard medical fine-tuning can compromise quantitative reasoning. The extent of degradation also depends on the base model: with a similar fine-tuning strategy, Meditron3 built on Llama-3.1 loses substantially more performance than its Qwen2.5-based counterpart, highlighting the importance of base-model choice. MedGemma further illustrates task-specific trade-offs: arithmetic drops only slightly, but comparison and aggregation decline by 20–60 points relative to its base. In contrast, models fine-tuned with chain-of-thought supervision (UltraMedical and Huatuo-o1) improve over their base models on comparison and aggregation, suggesting that reasoning-trace supervision can preserve or even enhance numeracy during medical adaptation.

Finding 3: Significant performance degradation under variant-template contexts indicates format sensitivity. Performance shifts across context representations depend on task complexity and model category. Base models exhibit sizeable robustness gaps between fixed-template and note-style variant contexts (10–22%), while medical models show smaller but still meaningful declines (0.5–15%). We also observe a trade-off: Llama-3.1 improves on variant templates relative to fixed templates, whereas CoT-trained medical models (UltraMedical and Huatuo-o1) degrade on variant contexts, suggesting that CoT fine-tuning can boost task performance but reduce format robustness. Even retrieval, despite its high absolute accuracy, can drop by 2–43% under formatting changes, indicating sensitivity in basic numerical extraction. Comparison shows the greatest robustness variability, with some models suffering catastrophic drops exceeding 22 points on certain variants. The robustness gap between 3 models groups shows that medical LLMs tend to have a lower robustness gap than general LLMs, highlighting the potential impact of format more than the length of the context. Overall, these results suggest that current LLMs lack format-invariant numerical representations and often rely on surface-level cues, which can fail under realistic documentation variation. This raises a critical

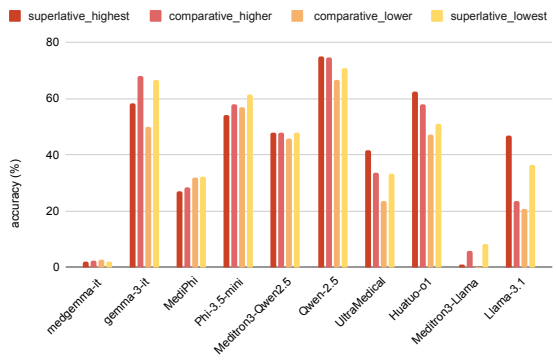

Figure 2: Fine-grained accuracy of comparison question types in fixed-template context.

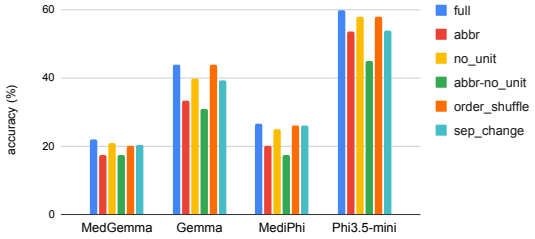

Figure 3: Aggregation accuracy when changing context template configuration. `full` denotes for the fixed template with full text and unit. `abbr` denotes for using abbreviations such as bp, hr, rr, o2sat. `sep_change` denotes for changing seperator, such as comma, colon.

concern for real-world clinical deployment.

4.2 Analysis

Comparison error analysis by question type. To better understand failures on comparison, we report accuracy by question type for both comparative and superlative queries: `highest`, `higher`, `lower`, and `lowest` (see Section 3.2.3 for definitions). Figure 2 reports results on the fixed-template context. A consistent pattern emerges: *lower comparative* questions are the most challenging for nearly all models, while the other types show larger model-to-model variation. One plausible reason is that `lower` comparatives require jointly tracking temporal order and applying a direction-sensitive threshold (e.g., finding the first time a vital drops below a value), which is prone to errors in inequality direction and record selection. In contrast, superlative questions can sometimes be answered by identifying a single extreme value, which may align better with common prompting heuristics. Overall, this breakdown suggests that comparison failures are not uniform, and that threshold-based "lower-than" reasoning is a key bottleneck even when models perform well on retrieval and arithmetic.

Aggregation robustness under note-style formatting variations. To investigate the performance drop under note-style variant contexts, we analyze aggregation accuracy under common documentation variations observed in patient notes. Using templates extracted in Section 3.1.3, we categorize prevalent surface-form changes, including abbreviations (e.g., BP, HR, RR, O_2 sat), separators (e.g., comma, colon, semicolon), and the presence or absence of units (e.g., "BP 120/70 mmHg" vs. "BP 120/70"). The distribution of these configurations is summarized in Appendix A.3 (Figure 4).

Figure 3 reports aggregation accuracy under controlled variants of the fixed-template context, including abbreviation only, unit removal only, and their combination. Results show that abbreviations and missing units are the primary drivers of errors. A plausible explanation is that abbreviations weaken the lexical cues that bind each number to the correct vital-sign attribute, while missing units remove an additional signal for interpreting values and disambiguating fields. When both factors co-occur, errors become more frequent, suggesting that models rely on redundant surface cues and degrade when multiple cues are removed simultaneously. In contrast, changes in attribute order or separator choice have a negligible effect, indicating that models are less sensitive to shallow punctuation changes than to information that directly supports value interpretation and field assignment.

Template substitution fidelity check. To ensure that our note-style variants preserve the intended measurements, we validate the extraction-and-substitution pipeline used to instantiate templates. We apply extraction matching to align vital-sign spans and values across variant representations, and *manually* audit 100 randomly sampled outputs for semantic correctness. Specifically, we verify that each substituted vital-sign value corresponds to the correct attribute in the underlying structured record, that units (when present) are consistent with the intended field, and that no attribute is dropped or duplicated after substitution. This audit achieves **96% correctness**, indicating that the constructed note-style variants largely reflect the underlying vital-sign records. Therefore, the performance differences observed across context representations are more likely attributable to representational variation rather than template corrup-

tion or substitution errors.

5 Conclusion

In this paper, we introduce ClinicNumRobBench, a benchmark that measures clinical numeracy correctness and robustness to documentation format. Built from longitudinal MIMIC-IV records with programmatic ground truth, ClinicNumRobBench evaluates four core numerical operations (retrieval, arithmetic, comparison, and aggregation) under three semantically equivalent context representations. Experiments on 17 LLMs reveal three limitations: comparison and aggregation remain challenging, especially for lower comparative queries; medical fine-tuning can reduce numeracy relative to base models, particularly for comparison; and note-style variation, especially abbreviations and missing units, can substantially degrade performance. Overall, ClinicNumRobBench provides a controlled testbed for diagnosing format sensitivity in clinical numerical reasoning and tracking progress toward more reliable numeracy for safer clinical deployment.

6 Discussion

Our systematic analysis yields actionable insights for both medical model developers and clinical practitioners. For developers, we identify four critical strategies to enhance model performance: (1) strategic selection of foundational architectures with robust numerical reasoning capabilities, (2) enrichment of training corpora with non-standard clinical formats and abbreviations, (3) augmentation of underrepresented task categories such as comparative reasoning instances, and (4) integration of chain-of-thought supervision to strengthen domain expertise and preserve quantitative reasoning abilities simultaneously. For clinicians, our findings underscore the importance of providing complete contextual information, particularly by including explicit units of measurement and avoiding the concurrent use of abbreviations with unit omissions, which substantially impairs model performance. Additionally, clinicians should exercise heightened caution when deploying these systems for numerical comparison and data aggregation tasks, where performance degradation is most pronounced.

Ethical Statement

The data utilized in this study does not contain any personally identifiable information or offensive content. All data is in English and originates from multiple sources, each distributed under different licenses. The Open Patients dataset (Khandekar et al., 2024) is sourced from multiple clinical open access corpora, including the Text Retrieval Conference (TREC) Clinical Decision Support/-Clinical Trials tracks, MedQA-USMLE, and PMC-Patients. MIMIC-IV (Johnson et al., 2023b) and MIMIC-IV-ED (Johnson et al., 2023a, 2024) are publicly available under PhysioNet credentialed access at https://doi.org/10.13026/kpb9-mt58 and https://doi.org/10.13026/5ntk-km72, respectively. A condition of releasing a derivative of the MIMIC datasets it to retain the same license. Therefore, the data is submitted and published under PhysioNet credentialed access.

We used AI assistants (manus, Claude, PaperReview) to refine Figure 1, enhance language, and review the manuscript.

Limitations

While our evaluation provides robust insights into LLM capabilities in medical numeracy, there are two primary limitations, which are potential for future investigation. First, our use of question templates enables systematic evaluation of robustness in context understanding and reasoning through controlled variations, ensuring reproducibility and isolated assessment of contextual factors affecting performance. However, clinical practice involves stakeholders expressing semantically equivalent queries through unconstrained natural phrasings. Assessing robustness using naturally expressed questions, such as those generated through LLM paraphrasing, LLM generation, or collected from practitioners, would advance understanding of LLM capabilities in capturing the full linguistic variability inherent in clinical settings. Second, we strategically focused on vital signs and related formulas (e.g., shock index, mean arterial pressure), selecting well-established medical concepts present in both structured data (EHRs) and unstructured text (patient notes) that are likely well-represented in LLM training data. This foundation enables confident conclusions about numerical reasoning on core medical concepts. However, the medical knowledge landscape encompasses numerous specialized domains that may be underrepresented in current models, potentially affecting retrieval and reasoning performance. Systematically extending our evaluation framework to these diverse knowl-

edge domains would provide comprehensive understanding of LLM capabilities and limitations across the medical knowledge spectrum, particularly regarding numerical reasoning proficiency.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 Technical Report. *arXiv preprint arXiv:2412.08905*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Pulkit Arora, Akbar Karimi, and Lucie Flek. 2025. Exploring Robustness of LLMs to Paraphrasing Based on Sociodemographic Factors.

Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, et al. 2025. MedHELM: Holistic Evaluation of Large Language Models for Medical Tasks. *arXiv preprint arXiv:2505.23802*.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025a. Benchmarking Large Language Models on Answering and Explaining Challenging Medical Questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599, Albuquerque, New Mexico. Association for Computational Linguistics.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024a. HuatuoGPT-o1, Towards Medical Complex Reasoning with LLMs. *arXiv preprint arXiv:2412.18925*.

Sijia Chen, Xiaomin Li, Mengxue Zhang, Eric Hanchen Jiang, Qin Zeng, and Chen-Hsiang Yu. 2025b. CARES: Comprehensive Evaluation of Safety and Adversarial Robustness in Medical LLMs. *ArXiv*, abs/2505.11413.

Zeming Chen, Angelika Romanou, Antoine Bonnet, Alejandro Hernández-Cano, Badr Alkhamissi, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, et al. 2024b. MEDITRON: Open Medical Foundation Models Adapted for Clinical Practice.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*.

Jean-Philippe Corbeil, Amin Dada, Jean-Michel Attendu, Asma Ben Abacha, Alessandro Sordoni, Lucas Caccia, François Beaulieu, Thomas Lin, Jens Kleesiek, and Paul Vozila. 2025. A Modular Approach for Clinical SLMs Driven by Synthetic Data with Pre-Instruction Tuning, Model Merging, and Clinical-Tasks Alignment. *arXiv preprint arXiv:2505.10717*.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 Herd of Models. *arXiv e-prints*, pages arXiv–2407.

Abinitha Gourabathina, Walter Gerych, Eileen Pan, and Marzyeh Ghassemi. 2025. The Medium is the Message: How Non-Clinical Information Shapes Clinical Decisions in LLMs. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 1805–1828.

Pierre Guiraud. 1959. Problèmes et méthodes de la statistique linguistique.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. *NeurIPS*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. *arXiv preprint arXiv:2009.13081*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

A Johnson, L Bulgarelli, T Pollard, B Gow, B Moody, S Horng, LA Celi, and R Mark. 2024. MIMIC-IV (Version 3.1). PhysioNet. RRID: SCR_007345.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Leo Anthony Celi, Roger Mark, and Steven Horng. 2023a. MIMIC-IV-ED v2.2. *PhysioNet*.

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

Nikhil Khandekar, Qiao Jin, Guangzhi Xiong, Soren Dunn, Serina Applebaum, Zain Anwar, Maame Sarfo-Gyamfi, Conrad Safranek, Abid Anwar, Andrew Zhang, et al. 2024. MedCalc-Bench: Evaluating

Large Language Models for Medical Calculations. *Advances in Neural Information Processing Systems*, 37:84730–84745.

Haoyang Li, Xuejia Chen, Zhanchao Xu, Darian Li, Nicole Hu, Fei Teng, Yiming Li, Luyu Qiu, Chen Jason Zhang, Li Qing, et al. 2025. Exposing Numeracy Gaps: A Benchmark to Evaluate Fundamental Numerical Abilities in Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20004–20026.

Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. 2024. GSM-Plus: A Comprehensive Benchmark for Evaluating the Robustness of LLMs as Mathematical Problem Solvers. *ArXiv*, abs/2402.19255.

Zefei Long, Zhenbiao Cao, Wei Chen, and Zhongyu Wei. 2025. EMGLLM: Data-to-Text Alignment for Electromyogram Diagnosis Generation with Medical Numerical Data Encoding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20470–20480.

Rahmad Mahendra, Damiano Spina, Lawrence Cavedon, and Karin Verspoor. 2024. Do numbers matter? Types and prevalence of numbers in clinical texts. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 409–415.

Rahmad Mahendra, Damiano Spina, Lawrence Cavedon, and Karin Verspoor. 2025. Evaluating Numeracy of Language Models as a Natural Language Inference Task. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8336–8361.

David Malvern, Brian Richards, Ngoni Chipere, and Pilar Durán. 2004. *Lexical Diversity and Language Development*. Springer.

Philip M McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Shreshth Mehrotra, Raghavendra P, Balraj Prajesh, Hrishikesh Kambale, and Puspita Majumdar. 2025. Unmasking Bias in Financial AI: A Robust Framework for Evaluating and Mitigating Hidden Biases in LLMs. *Proceedings of the 6th ACM International Conference on AI in Finance*.

Robert Osazuwa Ness, Katie Matton, Hayden S. Helm, Sheng Zhang, Junaid Bajwa, Carey E. Priebe, and Eric Horvitz. 2024. MedFuzz: Exploring the Robustness of Large Language Models in Medical Question Answering. *ArXiv*, abs/2406.06573.

Tuan Dung Nguyen, Thanh Trung Huynh, Minh Hieu Phan, Quoc Viet Hung Nguyen, and Phi Le Nguyen. 2024. CARER-ClinicAl Reasoning-Enhanced Representation for Temporal Health Risk Prediction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10392–10407.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron C Wallace. 2018. A Corpus with Multi-Level Annotations of Patients, Interventions and Outcomes to Support Language Processing for Medical Literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207.

Juliana Damasio Oliveira, Henrique DP Santos, Ana Helena DPS Ulbrich, Julia Colleoni Couto, Marcelo Arocha, Joaquim Santos, Manuela Martins Costa, Daniela Faccio, Fabio O Tabalipa, and Rodrigo F Nogueira. 2025. Development and evaluation of a clinical note summarization system using large language models. *Communications Medicine*, 5(1):376.

Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, Steven Bedrick, and William R Hersh. 2022. Overview of the TREC 2022 Clinical Trials Track. In *TREC*.

Russell L. Rothman, Victor M. Montori, Andrea L Cherrington, and Michael Pignone. 2008. Perspective: The Role of Numeracy in Health Care. *Journal of Health Communication*, 13:583 – 595.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. 2025. MedGemma Technical Report. *arXiv preprint arXiv:2507.05201*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Jun-Mei Song, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *ArXiv*, abs/2402.03300.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, et al. 2024. UltraMedical: Building Specialized Generalists in Biomedicine. *Advances in Neural Information Processing Systems*, 37:26045–26081.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate Before Use: Improving Few-Shot Performance of Language Models.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs. In *Conference on Empirical Methods in Natural Language Processing*.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. MedXpertQA: Benchmarking Expert-Level Medical Reasoning and Understanding.

A Additional Results

A.1 Arithmetic

Table 4 reports accuracy on arithmetic sub-tasks with increasing computational depth, including 1-step, 2-step, and 3-step calculations. Across models, accuracy generally decreases as the number of required operations increases, indicating that multi-step arithmetic remains substantially harder than single-step computation. Models trained or fine-tuned with reasoning-focused data show smaller performance gaps across depths, suggesting greater robustness to increasing computational complexity. Overall, these results highlight the importance of reasoning-oriented supervision for maintaining stable performance as arithmetic requires more steps.

Model	1-step	2-steps	3-steps
MedGemma	69.50	73.50	63.50
MediPhi	21.00	14.50	14.00
Meditron3-Qwen2.5	82.00	77.00	61.50
Meditron3-Llama	56.00	45.50	38.00
UltraMedical	51.50	38.50	38.00
Huatuo-o1	51.00	51.50	54.00
Gemma	79.00	72.00	68.00
Phi-3.5-mini	86.50	60.00	63.00
Qwen-2.5	90.50	78.50	78.00
llama-3.1	74.00	70.00	55.00
DeepSeek-R1-Distill	65.00	64.50	27.50
Qwen3	95.50	84.50	68.00
GPT-4.1-mini	96.50	92.50	88.50
Qwen3-reasoning	96.00	94.00	93.00

Table 4: Accuracy of Arithmetic sub-tasks.

A.2 Comparison

Table 6 reports detailed accuracy for the comparison task across the three context representations.

LLM	Model Implementation
medgemma-it	google/medgemma-4b-it
gemma-3-it	google/gemma-3-4b-it
MediPhi	microsoft/MediPhi-Instruct
Phi-3.5-mini	microsoft/Phi-3.5-mini-instruct
Meditron3-Qwen2.5	OpenMeditron/Meditron3-Qwen2.5-7B
Meditron3-Llama	OpenMeditron/Meditron3-8B
UltraMedical	TsinghuaC3I/Llama-3.1-8B-UltraMedical
Huatuo-o1	FreedomIntelligence/HuatuoGPT-o1-8B
Qwen-2.5	Qwen/Qwen2.5-7B-Instruct
llama-3.1	meta-llama/Llama-3.1-8B-Instruct
Qwen3	Qwen/Qwen3-8B
DeepSeek-R1-Distill-Llama-8B	unsloth/DeepSeek-R1-Distill-Llama-8B
GPT-4.1-mini	OpenAI API

Table 5: Details of used LLMs.

A.3 Template Variations

Table 9 reports accuracy under template variations that reflect common formatting patterns in clinical documentation. While the standardized template uses full attribute names, explicit units, and space separators, real clinical notes exhibit substantial variation. For example, attributes are often abbreviated (e.g., `heart_rate` as HR/hr and `oxygen_saturation` as O2sat/SpO2), units may be omitted, and separators such as commas or colons are frequently used. These variations capture realistic note formats and enable systematic assessment of model robustness to documentation inconsistencies.

B Benchmark Comparisons

We compare ClinicNumRobBench to existing benchmarks in Table 7. Table 2 provides a detailed comparison with MedCalc-Bench, which also evaluates numerical reasoning but focuses primarily on arithmetic computation.

C Lexical Diversity

To quantify the linguistic variation across our context representations, we measure lexical diversity for each format and compare our most realistic setting against MedCalc-Bench (Khandekar et al., 2024). Following standard practice in corpus linguistics, we use the root type–token ratio (Root TTR) (Guiraud, 1959):

$$\text{Root\_TTR} = \frac{\text{Number of unique tokens}}{\sqrt{\text{Total number of tokens}}}. \quad (1)$$

Higher Root TTR indicates greater vocabulary richness and surface-form variability.

Internal comparison across context formats. Tables 11 and 12 summarize internal and external comparisons, respectively. Internally, the note-style variant representation (Var-Tem.) exhibits substantially higher lexical diversity, with Root TTR ranging from 12.56 to 12.87, compared to structured (4.93–5.49) and fixed-template (4.38–5.51) representations. This makes Var-Tem. about two to three times more lexically diverse than the two controlled baselines, reflecting the note-derived templates used in Var-Tem., including abbreviations, unit omissions, and formatting differences that mirror real documentation practices.

Model	Struct.				Fixed				Variant			
	sup_highest	com_higher	com_lower	sup_lowest	sup_highest	com_higher	com_lower	sup_lowest	sup_highest	com_higher	com_lower	sup_lowest
medgemma-it	3.12	1.68	1.39	0.00	2.08	2.52	2.78	2.08	2.08	0.00	2.78	2.52
gemma-3-it	61.46	69.75	55.56	63.54	58.33	68.07	50.00	66.67	46.88	46.88	30.56	36.13
MediPhi	29.17	30.25	33.33	26.04	27.08	28.57	31.94	32.29	28.12	23.96	18.06	14.29
Phi-3.5-mini	53.12	54.62	52.78	52.08	54.17	57.98	56.94	61.46	44.38	47.50	48.89	46.97
Meditron3-Qwen2.50	44.79	32.77	27.78	38.54	47.92	47.90	45.83	47.92	36.46	37.50	26.39	34.45
Qwen-2.50	77.08	70.59	58.33	60.42	75.00	74.79	66.67	70.83	56.25	57.29	54.17	45.38
UltraMedical	27.71	25.97	19.72	33.96	41.67	33.61	23.61	33.33	19.79	21.88	16.67	15.97
Huatuo-o1	52.08	52.94	33.33	55.21	62.50	57.98	47.22	51.04	32.29	23.96	25.00	29.41
Meditron3-Llama	4.17	0.84	1.39	3.12	1.04	5.88	0.00	8.33	5.21	4.17	2.78	0.84
Llama-3.1	25.62	15.04	15.56	17.29	46.88	23.53	20.83	36.46	30.62	34.79	19.17	17.73
DeepSeek-R1-Distill-Llama	61.46	73.95	59.72	57.29	60.42	62.18	52.78	66.67	37.71	43.96	39.44	48.57
Qwen3-NonReasoning	76.04	66.39	59.72	81.25	80.21	75.63	68.06	73.96	50.00	50.00	45.83	51.26
GPT-4.1-mini	89.58	99.16	100.00	92.71	90.62	99.16	100.00	90.62	84.38	84.38	84.72	79.83
Qwen3-Reasoning	89.58	99.16	97.22	92.71	90.62	97.48	100.00	93.75	72.92	81.25	79.17	76.47

Table 6: Accuracy of comparison sub-tasks. `sup_highest` denotes superlative comparisons identifying the highest value, `sup_lowest` denotes superlative comparisons identifying the lowest value, `com_higher` denotes comparative assessments of whether one value is higher than another, and `com_lower` denotes comparative assessments of whether one value is lower than another.

	Medical	Qual. Reasoning	Comput.	Non-MCQ	Num.	Granul.	Robust.
MedQA	✓	✓	✗	✗	✗	✗	✗
MedXpertQA	✓	✓	✗	✗	✗	✗	✗
MedBullets	✓	✓	✗	✗	✗	✗	✗
GSM8K	✗	✓	✓	✓	✗	✗	✗
MATH	✗	✓	✓	✓	✗	✗	✗
NumericBench	✗	✓	✓	✗	✓	✓	✗
MedCalc	✓	✓	✓	✓	✗	✗	✗
Ours	✓	✓	✓	✓	✓	✓	✓

Table 7: Comparison of clinical and numerical reasoning benchmarks for LLM evaluation. Medical: tasks for medical evaluation; Qualitative (Qual) Reasoning: dataset tests qualitative reasoning; Comput.: dataset requires computation (i.e., quantitative reasoning); Non-MCQ: questions which have a single answer and without the use of multiple choices; Numeracy (Num.): dataset for numeracy evaluation; Granularity (Granul.): dataset test fine-grained levels in particular problem; Robustness (Robust.): the problem space is modeled using several data views/structures.

Task Category	Measurement	Temporal	Ratio/Proportion	Range	Frequency	Ordinal	Formula/Math
Retrieval							
Retrieval	✓	✓					
Arithmetic							
1-step	✓	✓					✓
2-steps	✓	✓	✓			✓	✓
3-steps	✓	✓	✓	✓			✓
Comparison							
Comparative	✓	✓		✓			
Superlative	✓	✓		✓		✓	
Aggregation							
Aggregation	✓	✓		✓	✓	✓	✓

Table 8: Alignment of benchmark tasks with clinical numerical information types of Mahendra et al. (2024).

External comparison with MedCalc-Bench. Relative to MedCalc, our Var-Tem. contexts have Root TTR that is about two points higher (12.71 vs. 10.71) and HDD that is about 0.12 higher (0.85 vs. 0.73), indicating greater local vocabulary variation. By contrast, MedCalc-Bench attains higher MTLD (96.10 vs. 84.66), suggesting lexical diversity is maintained over longer passages. This divergence is expected because Root TTR and HDD emphasize local surface variation introduced at the record level, whereas MTLD captures sustained diversity across extended text and is influenced by MedCalc-Bench's longer narrative note style.

Overall, these metrics confirm that ClinicNumRobBench spans a controlled spectrum of linguistic complexity, from structured and fixed-template baselines to a note-style variant that better reflects real clinical documentation variability.

D The Prevalence of Numerical Information in Clinical Settings

Mahendra et al. (2024) documents the prevalence of numerical information in clinical data. To assess the impact of numerical content on model performance, we conduct extensive evaluation on

Model	full	abbr	no_unit	abbr-no_unit	order_shuffle	sep_change
MedGemma	21.87	17.41	20.90	17.41	19.90	20.40
Gemma	43.79	33.29	39.76	30.81	43.74	39.26
MediPhi	26.37	19.90	24.88	17.41	25.87	25.87
Phi3.5-mini	59.7	53.53	57.71	44.78	57.71	53.73
Meditron3-Qwen2.5	51.74	48.26	50.74	48.25	48.26	48.76
Qwen-2.5	58.21	50.25	55.72	50.72	55.72	55.22
Meditron3-Llama	16.92	17.41	17.91	14.43	16.42	16.42
UltraMedical	30.85	33.83	32.84	36.82	31.84	30.35
Huatuo-o1	35.32	30.85	32.84	32.84	36.82	34.33
llama-3.1	29.85	26.87	25.37	27.86	29.85	28.36
DeepSeek-R1-Distill	34.33	32.81	33.82	27.36	34.82	34.31
Qwen3	76.62	75.60	74.61	74.13	75.62	74.13
GPT-4.1-mini	82.09	80.82	81.76	79.15	81.91	81.14
Qwen3-reasoning	82.59	81.08	82.08	80.60	82.08	82.09

Table 9: Accuracy of the aggregation task when changing a configuration in the context template. `full` denotes for the fixed template with full text and unit. `abbr` denotes for using abbreviations such as bp, hr, rr, o2sat. `sep_change` denotes for changing seperator, such as comma, colon, semicolon.

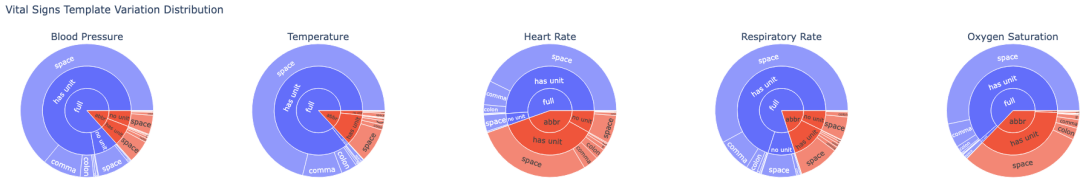

Figure 4: The distribution of template variations. `full` denotes full texting of vital sign name, such as *blood pressure*, opposite to `abbr` using abbreviations such as *bp*. The lightest circle denotes the separator between the vital sign name and its value used in the template

challenging benchmarks including MedXpertQA (Zuo et al., 2025) and MedBullets (Jin et al., 2020). Table 13 shows the prevalence of numerical data across these benchmarks, while Table 14 presents corresponding accuracy results. Our analysis reveals that numerical information appears in 73.51–100% of questions across challenging medical benchmarks and real-world data, yet model performance drops around 0.3% on number-containing questions compared to low overall accuracy, revealing a critical weakness in current LLMs and the need of attention on clinical numeracy.

E Prompts

The Zero-shot CoT Prompt

You are a helpful clinical assistant in <task>. Please think step-by-step to solve the question. Your final output should end with a JSON dict formatted as {"answer": <short_and_direct_answer_of_the_question>}.

Here is the task:
<question>

Let's think step-by-step to solve the question.

Figure 5: The zero-shot CoT prompt used for evaluation. <task> denotes the evaluating task, which is retrieval/arithmetic/comparison/aggregation.

Task	Subtask	ID	Question template	Reasoning focus
retrieval	retrieval	T1-6	`What was the {column} of the patient at {charttime}?`	Single-record lookup of a vital at a specific timestamp.
calculation	1step_add	T7	`Calculate the sum of systolic and diastolic blood pressure at {charttime}.`	A step in Mean Arterial Pressure (MAP)
calculation	1step_sub	T8	`Calculate the pulse pressure at {charttime}.`	Difference (SBP – DBP)
calculation	1step_mul	T9	`What is the value when multiplying the temperature at {charttime} by 1.8 round to 1 decimal places.`	Scalar multiplication (°C × 1.8)
calculation	1step_div	T10	`Calculate the Shock Index at {charttime}, round to 1 decimal places.`	Heart-rate / SBP ratio.
calculation	2step_add_div	T11	`Calculate the average of respiratory rate at the {case} 2 records, round to 1 decimal places.`	Mean of first or last two respiratory-rate entries.
calculation	2step_sub_div	T12	`Calculate a third of the pulse pressure at {charttime}, round to 1 decimal places.`	Pulse-pressure followed by division by three.
calculation	2step_add_mul	T13	`Convert temperature at {charttime} from Celsius to Fahrenheit, round to 1 decimal places.`	Full °C→°F conversion (×1.8 then +32).
calculation	2step_sub_mul	T14	`Calculate the percentage that heart rate at {charttime} exceeds baseline (70 bpm), round to 1 decimal places.`	Percent deviation from a fixed baseline.
calculation	3step_map	T15	`Calculate the Mean Arterial Pressure at {charttime}, round to 1 decimal places.`	MAP formula combining SBP and DBP.
calculation	3step_change	T16	`Calculate the percentage change in heart rate from {charttime_1} to {charttime_2}, round to 2 decimal places.`	Relative change across two timestamps.
comparison	comparative	T17-21	`What is the record datetime when {vital_parameter} first exceed {threshold}?`	Detect first crossing of clinician-defined thresholds.
comparison	comparative	T22-27	`What is the record datetime when {vital_parameter} first drop below {threshold}?`	Detect first crossing of clinician-defined thresholds.
comparison	superlative	T28-33	`What is the record datetime when {vital_parameter} is highest?`	Identify extremal measurement max timestamps.
comparison	superlative	T34-39	`What is the record datetime when {vital_parameter} is lowest?`	Identify extremal measurement min timestamps.
aggregation	aggregation	T40-42	`How many time the patient has {issue}? Return the number of records.`	Count events such as tachycardia, MAP > 100 mmHg, shock index > 0.7.

Table 10: Complete set of question templates used in the MIMIC-IV-ED Med Numeracy preprocessing pipeline.

Task	**Structured**	**Fixed-Tem.**	**Var-Tem.**
Retrieval	4.9268	4.3826	12.5604
Calculation	5.3545	5.5130	12.5724
Comparison	5.4663	4.6588	12.8695
Aggregation	5.4866	4.7530	12.8291

Table 11: The lexical diversity of questions in three context representations using Root TTR (Guiraud, 1959)

Metric	**MedCalc**	**Var-Tem.**
Root TTR (Guiraud, 1959)	10.7108	12.7079
HDD (Malvern et al., 2004)	0.7335	0.8487
MTLD (McCarthy and Jarvis, 2010)	96.0998	84.6647

Table 12: The lexical diversity of questions between MedCalc and our dataset

Benchmarks & Datasets	**%Num**
MedXpertQA (Zuo et al., 2025)	73.51
Medbullets (Chen et al., 2025a)	99.35
MedQA (Jin et al., 2020)	85.55
MedCalc-Bench (Khandekar et al., 2024)	96.73
PubMedQA* (Jin et al., 2019)	96.50
EBM-NLP* (Nye et al., 2018)	90.00
TREC-CDS* (Roberts et al., 2022)	100.00

Table 13: The ratio of samples containing numerical information in question contexts, articles, or patient notes. * denotes for the value from Mahendra et al. (2024).

Extraction Prompt

You are a medical documentation specialist. Your task is to extract raw vital sign text in given patient note which contains temperature, heart rate, respiratory rate, oxygen saturation, blood pressure. Text extracted must be a exact match from the patient note. Output format follow the JSON schema:

```
{
  "$defs": {
    "ValueObj": {
      "properties": {
        "text": {"type": ["string", "null"],
                 "description": "Raw text of the value"},
        "number": {"type": ["string", "null"],
                   "description": "Number of the value"},
        "unit": {"type": ["string", "null"],
                 "description": "Unit of the value"}
      },
      "required": ["text", "number", "unit"]
    }
  },
  "properties": {
    "temperature": {"$ref": "#/$defs/ValueObj"},
    "heart_rate": {"$ref": "#/$defs/ValueObj"},
    "respiratory_rate": {"$ref": "#/$defs/ValueObj"},
    "oxygen_saturation": {"$ref": "#/$defs/ValueObj"},
    "blood_pressure": {"$ref": "#/$defs/ValueObj"}
  },
  "required": ["temperature", "heart_rate", "respiratory_rate",
               "oxygen_saturation", "blood_pressure"]
}
```

Patient Note:
<patient_note>

Figure 6: Extraction Prompt used to extract a match from patient notes for mapping variant templates of the variant context representation.

	MedXpertQA	**MedBullets**
Full data	13.06	48.70
Number-only	12.72	48.69
%NumSamples	73.51%	99.35%

Table 14: The accuracy of 2 challenging clinical benchmarks. %NumSample denotes the ratio of samples containing at least 2 numbers as numerical information in question. Number-only denotes the accuracy of these samples.