MIMIC-SR-ICD11: A Narrative-Driven Benchmark for Disease Prediction with Large Language Models

Anonymous ACL submission

Abstract

Early disease diagnoses can dramatically improve patient outcomes by enabling timely interventions, yet traditional approaches rely on laboratory and imaging data that require clinical visits and incur significant costs and delays. In this study, we introduce MIMIC-SR-ICD11 (MIMIC Self-Report with ICD-11), a dataset that transforms EHR discharge notes from the MIMIC database into first-person patient narratives and standardizes every diagnoses using WHO ICD-11 codes. We benchmark three leading large language models on overall accuracy (Hit @1 and F1 variants), sensitivity to candidate list length and ordering, and robustness across diseases of varying prevalence. Our experiments show that simply shortening the candidate list does not yield proportional gains in accuracy, and F1 scores even fall below a random-guess baseline. By splitting diseases into ten frequency-based groups, we uncover an unexpected accuracy dip for the most common conditions. To explain this phenomenon, we introduce two lexical specificity metrics: disease frequency–medical vocabulary size (DF-MVS) and medical term exclusivity score (MTES). These metrics demonstrate that generic, nondistinctive terminology drives prediction bias. To support future advances, we release our dataset as a standardized benchmark for the development of specialized medical diagnostic models.

1 Introduction

Disease diagnosis has become a central pillar of modern healthcare, enabling early detection and timely intervention for acute conditions, while also guiding lifestyle adjustments and medication regimens to prevent or slow chronic diseases. It is particularly valuable in resource-limited environments and helps individuals without medical expertise avoid a long search for the right provider.

More recently, large language models (LLMs) have demonstrated strong performance on clinical question–answering benchmarks (Singhal et al., 2023, 2025). These models are typically fine-tuned on exam-style question–answer datasets designed for medical students (Jin et al., 2020; Pal et al., 2022; Jin et al., 2019), but their training regime does not directly translate to real-world diagnostic workflows, because these exam-style benchmarks present well-defined questions with fixed answer options, whereas real-world diagnosis involves interpreting ambiguous, multi-symptom narratives. Datasets designed for automatic diagnosis systems, such as DX (Wei et al., 2018) and DDXPlus (Tchango et al., 2022), predict a patient's underlying disease from categorical symptom indicators. However, this representation often obscures important clinical details; for example, reducing "severe, intermittent chest pain radiating to the left arm" to a simple present/absent flag loses key information about intensity and distribution. Moreover, because these collections are built for fixed-label classification, models trained on them cannot readily incorporate new symptoms or expand to additional disease categories beyond the original set. Su et al. (Su et al., 2024) introduced a dataset that uses patient-authored free-text symptom descriptions for automated disease prediction, but it is confined to Chinese data and leaves English self-reports unexplored.

To address these challenges, we introduce an English-language dataset, MIMIC-SR-ICD11, which converts EHR discharge notes into first-person patient self-reports and standardizes diagnoses using World Health Organization (WHO) ICD-11 terminology. We benchmark three leading LLMs (ChatGPT, Claude and Gemini) across three evaluation dimensions: overall predictive accuracy (Hit1 and Macro-/Micro-/Sample F1), sensitivity to candidate-list length and ordering, and robustness across diseases of varying prevalence. Our experiments reveal systematic positional bias in model decoding and a surprising performance decline for the most frequent conditions. To explain this counterintuitive result, we propose two lexical specificity metrics: disease frequency–medical vocabulary size (DF–MVS) and medical term exclusivity score (MTES), which together expose a vocabulary bottleneck in high-frequency disease descriptions. We publicly release our dataset and evaluation code to support further research on narrative-driven diagnostic models.

2 Related Work

In recent years, several datasets have been released to support the development of automatic diagnosis systems. SymCat (Peng et al., 2018) synthesizes records for 90 diseases and their associated symptoms by sampling according to disease–symptom co-occurrence probabilities. DX (Xu et al., 2019) and Muzhi (Wei et al., 2018) provide realistic, multi-turn doctor–patient dialogues that closely mirror actual clinical interactions. Building on this work, DDXPlus (Tchango et al., 2022) adds primary, differential, and final diagnoses alongside binary, categorical, and multiple-choice symptom representations. Although these benchmarks differ in scale, they all reduce symptoms to fixed options, which simplifies annotation and model training but sacrifices descriptive richness and obscures subtle clinical details. Moreover, by framing diagnosis purely as a classification task that maps a rigid symptom schema to a predetermined disease set, they are ill suited for evaluating a language model's generative abilities in tasks such as predicting diseases from free-text symptom descriptions.

To leverage the generative capabilities of large language models, several recent corpora incorporate free-text symptom descriptions to enrich diagnostic datasets. These resources abandon fixed-slot schemas and simple classification setups in favor of capturing authentic clinical narratives within generative modeling frameworks. MSdiagnosis (Hou et al., 2024) and CMEMR (Jia et al., 2025) combine multi-section free-text narratives, covering chief complaints, history of present illness, past medical history, physical examination findings, and laboratory results with structured labels for primary, differential, and final diagnoses. Haodf (Su et al., 2024) mines unstructured patient self-reports from the Haodf online platform to infer diagnoses directly, thereby reducing the time clinicians spend navigating fragmented symptom communications. Although these corpora significantly enhance the expressiveness of symptom narratives and support multi-step reasoning with generative models, they are all Chinese-language datasets. To date, no English-language corpus has been released that leverages real-world, patient-authored free-text narratives for automated diagnosis. Another related resource is ChatDoctor (Li et al., 2023), which consists of multi-turn patient–doctor consultation dialogues harvested from the HealthCareMagic[1] platform. This dataset captures patients' natural-language symptom descriptions alongside physicians' follow-up questions and diagnostic advice. Its design goal is to train medical dialogue systems capable of generating plausible diagnostic recommendations rather than performing precise disease prediction. Although the conversations include physicians' inferences about the patient's condition, these are expressed at a coarse level using colloquial disease names that are neither standardized nor mapped to a formal coding scheme (e.g., ICD-10), limiting the dataset's utility for evaluating fine-grained diagnostic accuracy.

3 Data Construction

To enable low-cost, early-stage disease screening, we focused our dataset on common conditions that together account for over 80 percent of real-world cases. We built the dataset on the latest MIMIC-IV (Johnson et al., 2024) and MIMIC-IV-Note (Johnson et al., 2022) releases. The construction pipeline consists of two steps. First, we extract each admission's diagnoses from MIMIC-IV and normalize them to ICD-11 terminology. Second, we retrieve free-text symptom descriptions from the MIMIC-IV-Note EMR notes and rewrite them as first-person patient self-reports. We release four variants that differ in the number of disease categories, allowing evaluation across models with varying capacity and deployment needs (see Table 1 for statistics).

3.1 Disease Normalization

Since diagnostic codes in MIMIC-IV originate from different versions of the International Classification of Diseases (ICD-9 and ICD-10), we standardized all disease labels to ICD-11 terminology for consistency. We used the official ICD-11 API provided by the WHO[2] to map diagnosis descriptions to ICD-11 terms.

[1] www.healthcaremagic.com

[2] https://icd.who.int/icdapi

Dataset	Number of Samples	Avg ± SD (Med, Min–Max) Tokens per Report	Avg ± SD (Med, Min–Max) Diseases per Sample	Number of Disease Classes
MIMIC-SR-ICD11_2000	272,579	159.5 ± 52.71 (152, 14 - 594)	2.81 ± 1.73 (2, 1 - 43)	98
MIMIC-SR-ICD11_1000	282,996	159.9 ± 53.03 (152, 14 - 594)	3.07 ± 1.91 (3, 1 - 57)	173
MIMIC-SR-ICD11_500	288,767	160.06 ± 53.15 (152, 14 - 594)	3.26 ± 2.03 (3, 1 - 69)	275
MIMIC-SR-ICD11_200	295,281	160.2 ± 53.25 (152, 14 - 594)	3.44 ± 2.14 (3, 1 - 80)	498

Table 1: Statistics of dataset variants, where data_N includes only diseases appearing at least N times. Tokens per report and diseases per sample are reported as Avg ± SD (Median, Min–Max).

During the mapping process, we applied a subphrase matching strategy to handle excessively long or detailed diagnosis descriptions that often include causal or symptomatic elaboration and therefore fail to match directly against the WHO API. Specifically, when a full diagnosis description did not return a result, we extracted all contiguous subphrases of at least 80 percent of the original length, sorted them from longest to shortest, and queried each in turn until we obtained a valid mapping. This approach increased the likelihood of finding a relevant standardized concept for complex or verbose descriptions. To ensure high-precision normalization and minimize noise, we retain only ICD-11 entities marked as `important=True` by the WHO API, which indicates strong mapping confidence. We also exclude non-leaf codes (those with child subcategories) to prevent semantic overlap among disease labels. This filtering enhances consistency in both model training and evaluation and reduces label ambiguity.

3.2 Patient's self-report generation

The MIMIC-IV-Note dataset provides de-identified free-text hospital records for each patient, which typically include a mix of symptom descriptions, examination results, medical history, and social background information. To derive patient-style narratives suitable for large language model reasoning, we utilized ChatGPT[3](`gpt-4o-mini`) developed by OpenAI to convert these clinical notes into first-person self-reports. During this transformation, we explicitly instructed the model to filter out all clinician-generated content such as physical examination findings, diagnostic test results, and professional assessments, and retain only the subjective symptom descriptions as if recounted by the patient. The generated self-reports are written in natural language using complete sentences and framed from the patient's perspective, which facilitates downstream disease diagnoses by aligning the input format with how patients typically describe their health concerns in real-world scenarios.

4 Experiment

To evaluate the diagnostic performance of state-of-the-art general-purpose LLMs, we conducted experiments with three widely used models: ChatGPT (`gpt-4o-mini`), Claude[4] (`claude-3-7-sonnet-20250219`), and Gemini[5] (`gemini-2.5-flash-preview-04-17`). Throughout our experiments, we employed the standardized prompt shown in Figure 1, which compels the models to select diagnoses exclusively from our candidate list. In addition to overall performance, we investigated how variations in the candidate disease list affect the quality of LLM prediction, focusing on changes in both the order and length of the candidate list. Finally, we analyzed the models' performance across diseases of different prevalence levels to better understand their strengths and limitations.

User input: Analyze the patient's description and identify the diseases that best match the reported symptoms, strictly selecting from the provided candidate list. Return only the names of the selected diseases, separated by semi colons. Do not include any reasoning or explanation. **Candidate list:** [plasma cell myeloma, unstable angina, nontoxic single thyroid nodule, ...]. **Patient self-report:** "I am a male patient who has been experiencing persistent fevers, night sweats, and profound fatigue over the past few weeks. I have also noted significant weight loss, estimating around 15 pounds in the last two months. My energy levels have drastically decreased, to the point where I find it difficult to engage in any physical activity, and I have been sleeping excessively, averaging most of the day during a recent trip. Additionally, I have a persistent cough that is mostly unproductive and occasionally feels aggravated. I have had episodes of diarrhea, which included a possible bloody component. Throughout this time, I have been monitoring my symptoms closely, and my condition feels as though it is deteriorating.
Claude output: fever; gastrointestinal tract haemorrhage; weight loss nos; septicaemia

Figure 1: Illustration of the standardized prompt format (instruction, candidate list, and patient self-report) provided to the LLMs, together with a sample output from Claude.

[3] https://openai.com/api/

[4] https://docs.anthropic.com/en/api/overview

[5] https://ai.google.dev/gemini-api/docs/api-versions

Model	Hit@1	Macro F1	Micro F1	Sample F1	Avg Pred Len	Valid Pred Rate
ChatGPT	0.2077 ± 0.0040	0.1727 ± 0.0038	0.1806 ± 0.0069	0.1630 ± 0.0070	2.7033 ± 0.0153	0.995 ± 0.0010
Gemini	0.1887 ± 0.0091	0.1510 ± 0.0034	0.1637 ± 0.0076	0.1549 ± 0.0092	3.9533 ± 0.0764	0.9981 ± 0.0004
Claude	0.2080 ± 0.0036	0.1492 ± 0.0021	0.1632 ± 0.0033	0.1546 ± 0.0054	3.5000 ± 0.0400	0.9603 ± 0.0040

Table 2: Performance comparison on disease prediction. "Avg Pred Len" denotes the average number of diseases predicted per sample. "Valid Pred Rate" indicates the proportion of model outputs that fall within the predefined candidate disease list, indicating how often the outputs conform to the candidate list.

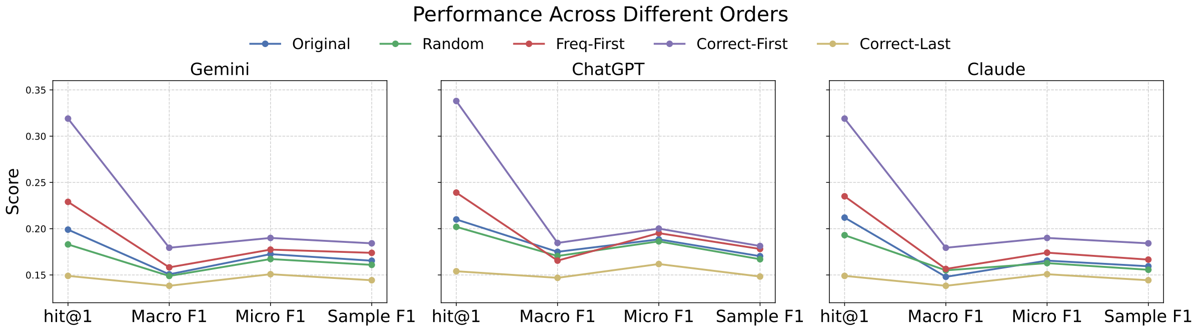

Figure 2: Comparison of model performance under different candidate disease orders.

4.1 Overall Performance

We split MIMIC-SR-ICD11_2000 into 80% training, 10% validation, and 10% test sets. From the test set, we drew three random subsamples of 1,000 records each. We evaluated our models on each subsample and then calculated the mean performance and its standard deviation across the three runs. The result is shown in Table 2.

Standard deviations for all metrics remain below 0.01, indicating high consistency across the three independent 1,000-sample. In terms of Hit@1, both ChatGPT and Claude achieve approximately 0.208, comfortably ahead of Gemini's 0.189. ChatGPT also leads on Macro, Micro, and Sample F1 scores. We attribute this to its more conservative strategy, favoring high-confidence diagnoses and producing shorter prediction lists. In contrast, Gemini and Claude generate longer lists to maximize recall, but this comes at the expense of F1 score. Finally, ChatGPT and Gemini strictly follow the candidate list provided, each with a Valid Pred Rate greater than 99%, while Claude's rate is slightly lower at 96%, indicating occasional out-of-list predictions.

4.2 Disease Order Effect

In this section, we investigate the robustness and stability of LLMs performance between different disease orders in the candidate list. Specifically, we design five schemes: 1) Origin is a random order but fixed for every sample prediction; 2) Random shuffles the candidate list for each sample; 3) Freq-First follows the real disease-frequency distribution (highest to lowest; 4) Correct-First places the ground-truth label at the front of the list; 5) Correct-Last places the ground-truth label at the end of the list. The result is shown in Figure 2.

As illustrated in the figure 2, varying the candidate list order produces a consistent effect across all LLMs: Correct-First ordering reliably achieves the highest performance, while Correct-Last ordering consistently produces the lowest. The superior performance of the Correct-First ordering can be largely attributed to the positional bias inherent in transformer-based LLMs. These models assign disproportionately higher attention weights to tokens that appear early in the input sequence, making them more likely to influence the output distribution. Comparing across LLMs, ChatGPT experiences the greatest boost under Correct-First ordering, with Hit@1 rising from around 0.2 to over 0.34. Gemini and Claude also benefit, each gaining about ten points in Hit@1, increasing from roughly 0.22 to around 0.32. This indicates that ChatGPT may be the most susceptible to positional cues but also the most capable of leveraging them when the correct label is highlighted. In contrast, Claude's performance drop under Correct-Last is the steepest among the three, confirming that it relies more heavily on early-positioned tokens.

Beyond these extreme cases, we also observe that ordering by global frequency ("Freq-First")

Model	Candidate List	Hit@1	Macro F1	Micro F1	Sample F1	Avg Pred Len	Valid Pred Rate
ChatGPT	Long List	0.210	0.1750	0.1885	0.1703	2.70	0.9963
	Short List	0.200	0.1738	0.1866	0.1673	3.06	0.9311
	Δ	-4.76%	-0.69%	-1.01%	-1.76%	–	-6.54%
Gemini	Long List	0.199	0.1507	0.1725	0.1654	3.97	0.9977
	Short List	0.197	0.1703	0.1899	0.1750	4.71	0.9518
	Δ	-1.01%	+13.01%	+10.09%	+5.80%	–	-4.60%
Claude	Long List	0.199	0.1507	0.1725	0.1654	3.97	0.9977
	Short List	0.205	0.1761	0.1943	0.1807	4.17	0.9451
	Δ	+3.02%	+16.85%	+12.64%	+9.25%	–	-5.27%

Table 3: Comparison between using long candidate lists and short candidate lists for disease prediction. Δ row shows the percentage change from the long list to the short list for each metric.

consistently improves over both the fixed original order and the per-sample random shuffle. This suggests a lightweight heuristic for boosting performance: listing candidates in decreasing prevalence can give the model a small but reliable edge, even without access to the true label. Notably, the gain from Freq-First is most pronounced for Hit@1, where accuracy rises by roughly 2–3 points all models.

Finally, we see that F1 metrics are relatively stable across ordering schemes compared to Hit@1. The divergence between the best and worst ordering for Macro and Sample F1 remains under 5 points for all models, whereas Hit@1 swings by over 15 percents. Because Hit@1 measures only the model's first choice, moving the true label out of the top slot causes Hit@1 to drop to zero. By contrast, F1 scores consider the full set of returned labels. Consequently, F1 metrics mitigate the impact of candidate list reordering, whereas Hit@1 remains highly sensitive to even slight shifts in label position.

4.3 Candidate Length Effect

We aim to test whether LLMs can distinguish the correct diagnoses when faced with high-confidence distractors in a compact candidate list. To this end, we first gather the false positive predictions from the models produced under the Freq-First and Correct-First orders. Because the models achieve their strongest performance under these two orders and the incorrect labels predicted by the models tend to assign high confidence score. We then insert the true diagnoses into this pool of confusable options. This process yields a shortlist that averages 10.45 diseases per case, allowing us to assess each model's ability to reject misleading choices. The results are summarized in Table 3

First, we observe that the Valid Pred Rate decreases for all models. The short candidate list contains each model's false positives, some of which fall outside the original candidate list. These invalid but confident distractors lower the Valid Pred Rate and force the models to cover more diseases than they did with the long list.

Second, shortening the candidate list yields clear gains for both Gemini and Claude. Gemini's F1 rises sharply simply because there are fewer distractors to sift through, but it still falls short of the other two models. However, the comparative performance of ChatGPT and Claude remains inconclusive. Claude outperforms ChatGPT on every metric, but part of this advantage may stem from its prediction length rather than from deeper diagnostic insight. Specifically, our shortlists contain an average of 10.45 diseases per case, with roughly three true labels. Randomly selecting four candidates from such a list yields a higher expected F1 score than selecting only three, since the extra pick increases the probability of covering a true label without substantially lowering precision. In practice, Claude outputs 4.17 labels on average, a tendency that artificially boosts its F1 even under near-random selection. ChatGPT, by contrast, returns fewer labels and therefore does not derive the same statistical benefit. This artifact explains why Claude appears to gain more from list reduction than ChatGPT, despite both models accessing the exact same information.

Finally, while these LLMs can easily exclude unlikely diseases, they struggle to distinguish among highly confusable conditions. Under a random-guess baseline on the short list, the expected sample F1 is approximately 0.30, yet all models perform well below this threshold. This gap highlights the need for stronger fine-grained diagnostic discrimination.

4.4 Performance in frequent and infrequent diseases

4.4.1 Counterintuitive Frequency Effects

We partitioned our 98 diseases into ten prevalence-based groups and evaluated models' accuracy within each group. Groups 1 through 9 each contain ten diseases sorted by descending frequency, while Group 10 comprises the eight rarest conditions. Figure 4 shows performance for each model under two candidate list ordering schemes: one arranged by global frequency (Freq-First) and the other randomly shuffled.

We find that LLMs do not exhibit a simple "frequency bias". Their Macro F1 scores across disease-frequency groups do not mirror the underlying data skew: rare conditions are not disproportionately more difficult to predict than common diseases. In fact, the decrease in Macro F1 from high- to low-frequency groups is much smaller than the corresponding drop in sample counts. This pattern is made more striking by a counterintuitive dip in performance on the very highest-frequency diseases (Group 1). Surprisingly, Group 1 yields lower F1 scores than Groups 2–4. Moreover, in classical imbalanced-label settings, models tend to overgenerate frequent labels, boosting recall at the expense of precision. In contrast, LLMs exhibit the opposite trend: for the highest-frequency group, precision markedly exceeds recall, whereas for the rarest groups, recall outstrips precision.

4.4.2 Mechanisms Behind Frequency Bias

Our first attempt employed a simple Document-Frequency–Vocabulary Size (DF–VS) metric, defined as

$$\text{DF-VS}_d = \frac{|V_d|}{\text{DF}_d} \tag{1}$$

where DF_d is the number of reports mentioning disease d and V_d the set of unique vocabularies used to describe it. Although low DF–VS correlates with poorer prediction (since a smaller, less diverse vocabulary makes discrimination harder), DF–VS alone does not fully account for the pronounced performance drop on the most frequent diseases. To uncover the additional factors behind this counterintuitive effect, we introduce two refined metrics, Frequency–Medical Vocabulary Size (DF–MVS) and Medical Term Exclusivity Score (MTES). Below, we define each metric formally.

Disease Frequency–Medical Vocabulary Size (DF–MVS) We denote by M_d the number of distinct medical terms appearing in all free-text reports for the disease d. The DF–MVS metric then captures the average term richness per document:

$$\text{DF-MVS}_d = \frac{M_d}{\text{DF}_d}. \tag{2}$$

Here, M_d is obtained by extracting all unique medical-entity mentions from reports of disease d, and DF_d is the total count of those reports. This metric captures both how often a disease appears and the breadth of specialized vocabulary used to describe it. Lower DF-MVS indicates that the description of a disease has fewer medical terms.

Medical Term Exclusivity Score (MTES) For each term $t \in V_d$, let $\text{freq}_{d'}(t)$ be the frequency of t in reports of the disease d', and let D be the total number of diseases. We first compute the average cross-disease frequency of t:

$$\overline{\text{freq}}_{-d}(t) = \frac{1}{D-1} \sum_{d' \neq d} \text{freq}_{d'}(t). \tag{3}$$

Then MTES is defined as the mean of the reciprocal of these averages (plus a small constant ϵ to avoid division by zero):

$$\text{MTES}_d = \frac{1}{|V_d|} \sum_{t \in V_d} \overline{\text{freq}}_{-d}(t). \tag{4}$$

Lower MTES values indicate that a disease's terms are rarely shared with others, reflecting higher descriptive specificity.

To calculate DF–MVS and MTES metrics, we employ spaCy's Bio-medical NER model (en_ner_bc5cdr_md) (Gu et al., 2017) to extract clinical entitie from free-text reports. This model has been fine-tuned on the BioCreative V CDR corpus (Li et al., 2016), which provides extensive, high-quality annotations for both disease and chemical mentions in biomedical literature. Compared to general-purpose NER systems, it offers superior precision and recall on medical terminology, ensuring robust recognition of both common and rare conditions.

As shown in Figure 4, although Group 1 diseases account for more than 40 percent of the dataset, they remain among the most difficult to predict for three closely linked reasons. First, Group 1 registers the lowest DF–VS values (Figure 4.a). This indicates that despite their high frequency, these

Group	1	2	3	4	5	6	7	8	9	10
Proportion (%)	43.49	17.83	9.64	6.98	5.50	4.45	3.82	3.24	2.93	2.13

Table 4: Sample Proportions in Each Disease-Prevalence Group (High-to-Low Frequency Order)

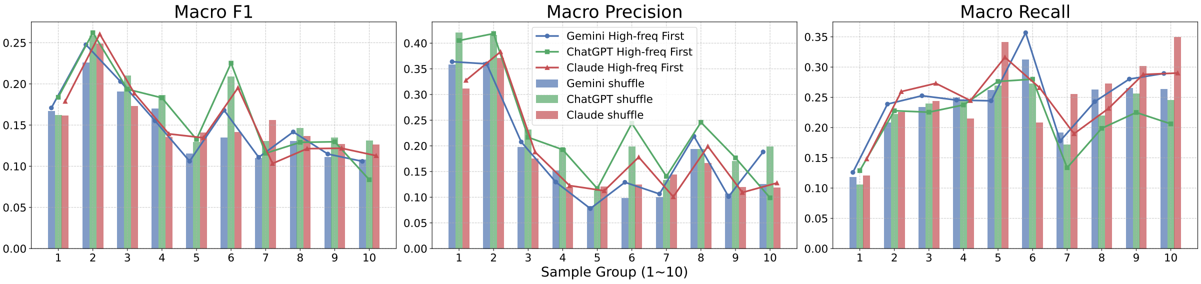

Figure 3: Performance across different frequency groups of diseases. Higher-frequency diseases tend to be predicted with greater accuracy, indicating a correlation between disease frequency and model performance.

diseases are described using a small, generic set of terms, which limits the richness of their lexical representation. Second, Group 1 also scores poorly on DF–MVS (Figure 4.b). In practice, this means that even though there is abundant text, very few of the tokens are truly informative medical words such as specific biomarkers, nuanced symptom qualifiers, or disease-defining expressions. As a result, narrative redundancy prevails with repeated mentions of terms such as pain or fever, depriving the model of the high-signal features it needs. Third, Group 1 has the lowest Medical Term Exclusivity Score (Figure 4.c). The few medical terms that appear are common across many conditions, so they carry little mutual information with any single diagnosis. Together, these factors create a vocabulary bottleneck: a small, non-exclusive set of terms that produces a low signal-to-noise ratio and prevents the model from learning the distinctive patterns required for accurate prediction.

By contrast, mid-frequency diseases in Groups 2 occupy an optimal range. They appear often enough to accumulate varied phrasing yet remain uncommon enough to invoke specific medical descriptors, and this balance raises their DF–MVS while lowering their MTES. Rare diseases in Groups 7 to 10 suffer from small sample sizes but benefit from highly distinctive labels such as "alopecia areata" or "Wegener's granulomatosis", which help the model form clear decision boundaries. Therefore, mid-frequency and rare disease categories tend to become easier for the model to predict simply by adding more examples, because the language used to describe them already contains distinct, high-signal terms. In contrast, adding or reducing additional samples for very common diseases does little to help, since their narratives rely on broad, generic words that volume changing alone cannot diversify.

To break through this plateau for frequent conditions, we need to deliberately expand the linguistic palette in their documentation. Pulling synonym lists from established medical vocabularies can inject subtle variations in how symptoms and findings are named. Paraphrase-driven templates can recast identical clinical observations in different phrasings, teaching the model to recognize conceptually equivalent expressions. Enhanced entity recognition can spotlight less obvious but diagnostically relevant terms buried in the text. Together, these efforts would boost the uniqueness and richness of the textual features, giving the model the nuanced cues it needs to distinguish among common disease labels. In addition to enriching narrative descriptions, we must acknowledge that some conditions simply lack highly distinctive symptoms and are only definitively diagnosed through laboratory tests or imaging studies. For these diseases, model inputs should be augmented with structured findings such as key lab values, radiology impressions or vital signs. Finally, we can improve prediction safety and focus by adopting a severity-weighted candidate checklist. Rather than treating all disease classes equally, the system can estimate each condition's potential impact on patient outcomes (e.g., risk of organ failure or mortality) and elevate high-consequence diagnoses into the shortlist even when their textual signatures are subtle. This risk-aware

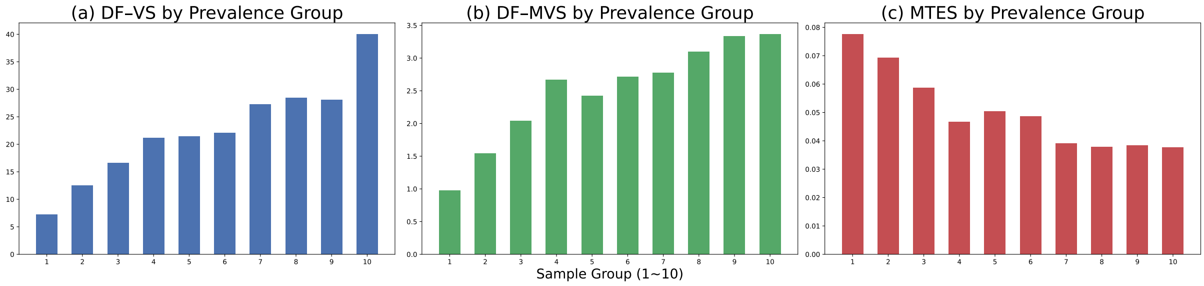

Figure 4: Analysis of disease prevalence and vocabulary specificity across prevalence groups. (a) Proportion of samples in each disease-frequency decile (Group 1 = most frequent, Group 10 = rarest). (b) Average Document-Frequency–Medical Vocabulary Specificity (DF–MVS) per group, illustrating that rarer diseases employ more distinctive medical terms. (c) Average Medical Term Exclusivity Score (MTES) per group, indicating how uniquely each group's terminology appears compared to the rest. A high DF–MVS coupled with a low MTES signifies that the group's medical vocabulary is both plentiful and uniquely characteristic.

ranking ensures that life-threatening conditions receive appropriate attention, balancing linguistic uncertainty with clinical urgency and further reducing the chance that a common but dangerous disease is overlooked.

5 Conclusion

In this work we have introduced a new benchmark for narrative-driven disease prediction. We convert MIMIC electronic health record notes into first-person patient self-reports and normalize all diagnoses to ICD-11 name. With the proposed MIMIC-SR-ICD11 dataset, we evaluate three leading large language models (ChatGPT, Claude and Gemini) on overall accuracy, sensitivity to candidate list ordering and length, and robustness across disease prevalence groups.

Our experiments reveal several important insights. First, LLMs performance depends strongly on how candidate diseases are presented. LLMs exhibit positional bias when disease lists are permuted and benefit from shorter, more focused candidate sets only under certain conditions. Second, common diseases prove surprisingly difficult to predict despite their data abundance. We show that this difficulty arises from a vocabulary bottleneck: high-frequency conditions are documented with generic and non-exclusive terms that offer low signal to the model. Mid-frequency and rare diseases by contrast gain more from increased sample size because their narratives already contain distinctive terminology.

6 Future Work

In the future, we identify three key directions for advancing narrative-driven disease prediction:

1. Incorporate Chain-of-Thought Inference. We will augment our dataset with explicit reasoning traces by prompting LLMs to generate intermediate "chain-of-thought" inference. This can guide models to attend to critical symptom-to-diagnoses links and improve interpretability, helping to surface latent clinical cues that raw text alone may obscure.

2. Train a Domain-Specialized Diagnostic LLM: We will fine-tune a compact model that has been pretrained on a large medical text corpus using our MIMIC-SR-ICD11 dataset. Training will include a reinforcement learning objective that rewards outputs matching our curated candidate list and penalizes any predictions outside this set. After fine-tuning, we will reassess the frequency bias to determine whether this approach improves prediction performance across both common and rare disease categories.

3. Enable hierarchical disease prediction with ICD-11 code We will leverage the built-in hierarchy of ICD-11 to produce predictions at multiple levels of granularity. For each specific code it predicts such as "Type 2 diabetes mellitus", it will also identify its parent categories (such as, "diabetes mellitus"). This layered output gives downstream applications the flexibility to adjust their decision scope. Downstream applications can then choose whether to act on broad disease classes for initial screening or on detailed leaf-level codes for precise diagnoses.

Together, these enhancements will deepen model reasoning, tailor performance to the diagnostic domain, and provide clinicians with both detailed and overview predictions across the disease hierarchy.

A License and Availability

All of our code and the MIMIC-SR-ICD11 dataset are released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. The full text of the CC BY 4.0 license is available at `https://creativecommons.org/licenses/by/4.0/`.

B Artifact Usage and Intended Use

B.1 Use of Existing Artifacts

We relied on several third-party resources, each with its own access and usage restrictions:

- MIMIC-IV and MIMIC-IV-Note are governed by PhysioNet's credentialed access agreement, which permits use for non-commercial, research-only purposes. All analysis in this paper was conducted under that agreement and in compliance with its requirement that no clinical or commercial deployment occur.
- WHO ICD-11 API is provided under the WHO open data policy. We used it strictly to normalize diagnoses in our dataset, in accordance with the API's terms for research and educational use.
- spaCy's en_ner_bc5cdr_md model is distributed under the Apache 2.0 license. Our extraction of medical entities from free-text reports adheres to that license, which allows both research and derivative work without additional restriction.

In each case, our usage matches the intended research-only context and does not violate any access conditions or licensing terms.

B.2 Intended Use of Created Artifacts

We release the MIMIC-SR-ICD11 dataset (and accompanying code) under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. This license permits unrestricted research and educational reuse, provided that appropriate credit is given. In keeping with PhysioNet's original agreement, we explicitly restrict MIMIC-SR-ICD11 to non-commercial, research-only applications. Users who wish to employ the dataset for clinical decision support or commercial purposes must first secure the necessary permissions from the data providers.

C Ethics

All human-subject data in this study (MIMIC-IV and MIMIC-IV-Note) are fully de-identified under HIPAA and were released under a PhysioNet Data Use Agreement (DUA) that mandates human-subjects training and forbids any attempt at re-identification. The original IRBs at Beth Israel Deaconess Medical Center and MIT approved the data release with a waiver of informed consent and determined that secondary analyses of these de-identified records are exempt from ongoing review. Our usage strictly adheres to these terms and remains within a research-only context. The patient self-reports we generate via ChatGPT are synthetic paraphrases of those de-identified notes and contain no additional private information, so no new consent is required.

D Acknowledgements

We used ChatGPT (gpt-4o-mini) as a writing assistant to help polish sentence structure and improve readability. All technical content, analyses, and conclusions were developed solely by the authors.

E Limitations

Despite its contributions, our work has several important limitations.

Synthetic self-reports Our patient narratives are generated by prompting ChatGPT to paraphrase de-identified discharge text. Although we carefully instructed the model to retain only subjective symptom language, this process may introduce artifacts, omit subtle nuances, or diverge from how real patients describe their own experiences. Future work should validate our synthetic reports against authentic patient-provided narratives.

ICD-11 mapping We rely on subphrase matching to map heterogeneous ICD-9/10 descriptions to ICD-11 concepts via the WHO API. While this increases coverage, it may still produce incorrect or overly broad mappings for very complex diagnoses. Mis-mappings could propagate noise into both training and evaluation.

References

Jinghang Gu, Fuqing Sun, Longhua Qian, and Guodong Zhou. 2017. Chemical-induced disease relation extraction via convolutional neural network. *Database*, 2017:bax024.

Ruihui Hou, Shencheng Chen, Yongqi Fan, Lifeng Zhu, Jing Sun, Jingping Liu, and Tong Ruan. 2024. Msdiagnosis: An emr-based dataset for clinical multi-step diagnosis. *arXiv preprint arXiv:2408.10039*.

Mingyi Jia, Junwen Duan, Yan Song, and Jianxin Wang. 2025. medIKAL: Integrating knowledge graphs as assistants of LLMs for enhanced clinical diagnosis on EMRs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9278–9298, Abu Dhabi, UAE. Association for Computational Linguistics.

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 W. 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.

Alistair E. W. Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger G. Mark. 2024. MIMIC-IV (version 3.1). Accessed: 2025-05-13.

Alistair E. W. Johnson, Tom J. Pollard, and Roger G. Mark. 2022. MIMIC-IV-Note: Deidentified free-text clinical notes (version 2.2).

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database (Oxford)*, 2016. Published May 9 2016; article baw068.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6):e40895.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning (CHIL)*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Yu-Shao Peng, Kai-Fu Tang, Hsuan-Tien Lin, and Edward Chang. 2018. Refuel: Exploring sparse features in deep reinforcement learning for fast disease diagnosis. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and 13 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180. Published 2023/08/01.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, and 16 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950. Published 2025/03/01.

Zhixiang Su, Yinan Zhang, Jiazheng Jing, Jie Xiao, and Zhiqi Shen. 2024. Enabling patient-side disease prediction via the integration of patient narratives. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 581–584, New York, NY, USA. Association for Computing Machinery.

Arsène Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. 2022. Ddxplus: a new dataset for automatic medical diagnosis. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuan-Jing Huang, Kam-Fai Wong, and Xiang Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.