# Beyond Prompting: What Information Matters for Clinical Diagnosis with Structured Inputs

**Anonymous EMNLP submission**

## Abstract

Large language models (LLMs) have shown promise in medical tasks, yet their ability to perform reliable clinical diagnosis from clinical records remains underexplored. While prior work often assumes prompting is sufficient, this study challenges that assumption by examining how information organization and supervised adaptation enable diagnosis under realistic admission-time settings. Clinical diagnosis is formulated as a prediction task using a leakage-reduced benchmark derived from MIMIC-IV. Across multiple LLM backbones, prompt-only methods perform poorly (2–13% EM), whereas supervised adaptation significantly improves performance (up to 75.44% EM). To better represent clinical evidence, a structured admission-time representation is proposed to organize records into clinically meaningful sections. Structured inputs are found to consistently outperform raw truncated notes, with further analysis suggesting that performance gains arise from preserving organized evidence rather than aggressive heuristic filtering. Controlled ablations reveal a clear hierarchy of diagnostic signals: narrative clinical history—specifically the History of Present Illness (HPI)—provides the dominant predictive signal, while objective findings offer complementary refinement. These findings suggest that LLM-based clinical diagnosis is best understood as a structured supervised learning problem where the organization of clinical evidence is critical.

**Keywords:** Clinical NLP, Admission-Time Diagnosis, Large Language Models, Supervised Adaptation, Clinical Reasoning

## 1 Introduction

Large language models (LLMs) have demonstrated strong performance across a wide range of natural language processing (NLP) tasks, including question answering, summarization, and dialogue (Brown et al., 2020). In the medical domain, recent work has explored LLMs for clinical decision support, medical question answering, and report generation (Jin et al., 2021; Singhal et al., 2023; Krishna et al., 2021; Nori et al., 2023; Tu et al., 2024). However, a fundamental question remains: *what information enables LLMs to perform clinical diagnosis, and can reliable diagnosis be achieved through prompting alone?*

Clinical diagnosis differs fundamentally from standard NLP tasks. Rather than relying on isolated facts or localized reasoning, diagnosis requires integrating heterogeneous evidence sources—including patient-reported symptoms, medical history, physical examination, laboratory findings, and imaging results—often under incomplete information and severe time constraints. In clinical practice, physicians frequently form early diagnostic hypotheses primarily from narrative history, particularly the History of Present Illness (HPI), while objective findings are used to refine nearby distinctions. Whether LLMs rely on similar information patterns, however, remains insufficiently understood.

Existing evaluations of medical LLMs often rely on medical exam questions (Liévin et al., 2024) or retrospective discharge summaries. In the latter case, substantial diagnostic information is often explicitly stated within the notes (Lehman, 2024), making it difficult to distinguish genuine diagnostic reasoning from information leakage or surface-level retrieval. Moreover, relatively little attention has been paid to the systematic organization and representation of clinical evidence for diagnostic prediction.

To address these limitations, this work formulates clinical diagnosis as a prediction task from admission-time records to coarse ICD-10 prefixes (e.g., the first three characters of ICD-10 codes). The task is evaluated using a leakage-reduced benchmark derived from MIMIC-IV (Johnson et al., 2023), a large-scale electronic health

record database. By systematically removing post-discharge and retrospective information, the task better approximates realistic early-stage diagnostic settings. This setup enables a controlled analysis of both information usage and representation structure in LLM-based diagnosis.

Under this setting, we systematically investigate how structured clinical evidence contributes to diagnostic performance. Figure 1 illustrates the proposed framework, which organizes clinical records into clinically meaningful sections and enables a controlled comparison between structured representations and raw truncated text. Controlled ablations further examine the relative contributions of narrative history, laboratory findings, imaging results, and physical examination findings.

Overall, this work makes four main contributions:

- **Benchmark Construction:** We introduce a leakage-reduced admission-time diagnosis benchmark derived from MIMIC-IV for controlled ICD-10 prediction.
- **Adaptation Analysis:** Systematic experiments demonstrate that reliable diagnosis depends substantially more on supervised adaptation than on prompting alone.
- **Structured Representation:** Structured clinical representations improve diagnostic prediction over raw truncated text by better preserving structured clinical evidence.
- **Information Hierarchy:** We identify a hierarchy of information usage in LLM-based diagnosis, where narrative history provides the dominant signal and objective findings contribute complementary refinement.

## 2 Related Work

**Medical LLMs and Clinical Diagnosis.** LLMs have demonstrated significant potential in medical NLP, spanning question answering (Jin et al., 2021; Singhal et al., 2023), clinical text generation (Krishna et al., 2021), and automated ICD coding (Dong et al., 2021). While benchmarks such as MedQA and Med-PaLM assess medical reasoning performance, other studies have focused on clinical prediction and coding from retrospective discharge summaries (Akkhawatthanakun et al., 2026; Rodrigues and Lopes, 2025). As discussed by Lehman (2024), such retrospective clinical documents are highly susceptible to information leakage, as notes authored after treatment completion often contain explicit diagnostic labels or definitive clinical summaries. Furthermore, most existing studies focus on prompting static benchmarks rather than direct inference from raw clinical records (Tu et al., 2024). In contrast, we study admission-time diagnosis under leakage-reduced settings, requiring models to infer diagnoses from noisy and incomplete evidence available during early patient encounters.

**Prompting, In-Context Learning, and Supervised Adaptation.** Prompting and in-context learning (ICL) are widely used to adapt LLMs to downstream tasks without parameter updates (Brown et al., 2020). In the medical domain, prior work has extensively explored zero-shot and few-shot prompting for clinical reasoning and medical QA (Liévin et al., 2024; Tu et al., 2024). Parameter-efficient fine-tuning (PEFT) methods, such as LoRA (Hu et al., 2022), further enable lightweight supervised adaptation of LLMs to specialized clinical tasks (Han et al., 2023). However, recent evidence suggests that prompting alone remains insufficient for reliable performance in complex, real-world clinical settings (Liévin et al., 2024). Unlike standardized medical exams, admission-time diagnosis requires integrating heterogeneous evidence from raw clinical records, which is difficult to capture through prompting alone (Fornasiere et al., 2024). Furthermore, specialized supervised adaptation often provides more stable performance on clinical prediction tasks (Tu et al., 2024; Han et al., 2023). Our results further support this observation: supervised adaptation substantially outperforms prompt-only methods in leakage-reduced admission-time diagnosis.

**Structured Clinical Representation and Input Organization.** Clinical NLP systems have traditionally treated electronic health records (EHRs) as flat, unstructured text sequences (Alsentzer et al., 2019). To better capture the functional organization of clinical narratives, researchers have developed specialized models for clinical note section classification and boundary detection (Zhou et al., 2023). Modular frameworks such as medspaCy further provide rule-based pipelines for sentence splitting and section detection (Velamala et al., 2025).

However, prior work has primarily focused on segmentation quality or performance on localized extraction tasks. Relatively little attention has been paid to the diagnostic utility of differ-

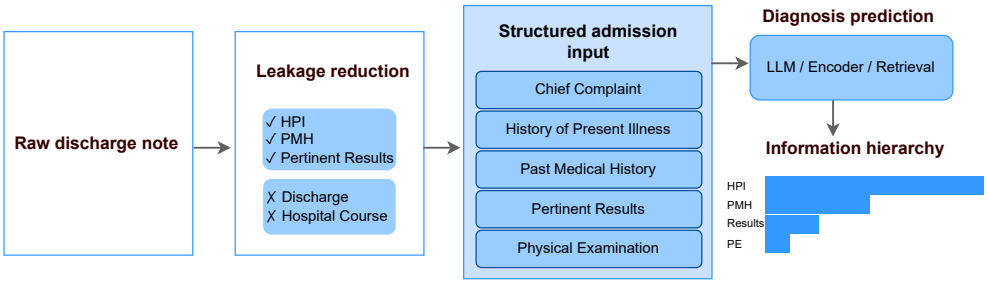


Figure 1: Overview of the admission-time diagnosis framework. Retrospective, discharge-related sections are removed to mitigate information leakage. The remaining clinical evidence is organized into clinically meaningful sections before diagnostic prediction. Controlled analyses further reveal a hierarchy of information usage in which narrative history contributes the dominant diagnostic signal.

ent clinical sections, or how section organization impacts admission-time diagnosis. In contrast, our work systematically evaluates structured clinical representations and provides a controlled quantitative analysis of how distinct information sources—ranging from narrative patient history to objective findings—contribute to diagnostic performance.

**Information Usage and Diagnostic Reasoning Analysis.** Recent work has increasingly focused on how LLMs utilize input information through attribution methods, perturbation analysis, and retrieval-based evaluation (Liu et al., 2024). In the medical domain, studies have primarily examined the general reasoning capabilities of foundation models and the effectiveness of medical chain-of-thought (CoT) prompting (Tu et al., 2024; Wang et al., 2022). However, as highlighted by Liévin et al. (2024), it remains unclear how models perform in real-world scenarios requiring expert domain knowledge beyond standardized benchmarks. Most existing analyses rely on qualitative inspection or overall performance metrics, rather than systematically quantifying how different clinical information sources contribute to diagnosis. Our work addresses this gap by providing a controlled quantitative analysis of information usage in admission-time diagnosis. Our results reveal a hierarchy of diagnostic signals, demonstrating that while narrative patient history provides the dominant signal, objective findings provide complementary refinement.

# 3 Method

## 3.1 Task Formulation

Clinical diagnosis is formulated as a supervised task from admission-time clinical text $x$ to a coarse ICD label $y$, represented as a three-character ICD-9/ICD-10 prefix. Restricting inputs to admission-time information reduces post-discharge leakage and better reflects realistic early-stage diagnostic settings (Lehman, 2024).

The label space consists of $|\mathcal{Y}| = 50$ frequent ICD prefixes, constructed by retaining the most common primary diagnoses prior to patient-level splitting. ICD-9 and ICD-10 codes are normalized into a unified three-character prefix space to ensure consistency across coding systems (Dong et al., 2021).

Prediction is implemented through conditional generation. The model produces a short continuation, from which the first valid three-character ICD token is extracted using a deterministic normalization rule. Exact match (EM) over normalized ICD prefixes is used as the primary evaluation metric, enabling consistent comparison across prompting, retrieval, and supervised settings. Additional relaxed-matching analyses further confirm that prompting gaps are not primarily explained by output-formatting artifacts.

This formulation serves as a controlled proxy for coarse diagnosis. Using coarse ICD prefixes reduces sparsity and label ambiguity while preserving clinically meaningful disease categories, but is not intended to capture full clinical coding complexity, fine-grained diagnostic reasoning, or replace real-world clinical diagnosis.

### 3.2 Leakage-Reduced Input Construction

Clinical notes are processed using a deterministic section-based pipeline that retains admission-time information while removing retrospective or discharge-related content. The pipeline is designed to approximate information available during early-stage diagnostic assessment while minimizing retrospective leakage from discharge-oriented documentation (Lehman, 2024).

Five clinically motivated sections are retained: Chief Complaint (CC), History of Present Illness (HPI), Past Medical History (PMH), Pertinent Results, and Physical Examination (PE). These sections are concatenated in a fixed order:

$$\text{CC} \rightarrow \text{HPI} \rightarrow \text{PMH} \rightarrow \text{Results} \rightarrow \text{PE}.$$

This structure reflects the typical flow of clinical reasoning, where narrative history precedes objective findings. It also maintains semantically distinct evidence sources within a coherent representation. Separating heterogeneous evidence sources into clinically coherent components also enables controlled analysis of how different forms of clinical information contribute to diagnosis (Fornasiere et al., 2024).

The Results section can optionally be compressed using heuristic filtering to retain concise high-signal findings (e.g., abnormal laboratory values or short imaging summaries). Controlled ablations evaluate the impact of this compression on diagnostic performance.

### 3.3 Models

To evaluate the role of supervision and input organization under a unified task formulation, we compare multiple adaptation paradigms, including prompting, retrieval, encoder classification, and supervised LLM fine-tuning.

Qwen2.5-7B-Instruct (Team, 2024) is used as the primary backbone. Three settings are evaluated under a unified task formulation:

1. zero-shot prompting,
2. in-context learning (ICL) (Brown et al., 2020),
3. supervised fine-tuning with LoRA (Hu et al., 2022).

To assess generality, LoRA is additionally applied to multiple instruction-tuned backbones under the same data and input setup.

### 3.4 Retrieval and Encoder Baselines

BM25+kNN retrieves top-$k$ similar training cases using sparse lexical matching (Robertson and Zaragoza, 2009) and predicts labels via majority voting. Dense kNN encodes inputs using frozen BioClinicalBERT (Alsentzer et al., 2019) representations and performs nearest-neighbor classification in embedding space.

The encoder baseline trains BioClinicalBERT with a linear classification head over the same label space, providing a discriminative comparison to generative LLMs.

### 3.5 Supervised Fine-Tuning

LoRA is used to fine-tune the LLM by injecting low-rank updates into attention layers while keeping backbone weights frozen. This enables efficient adaptation while preserving pretrained model parameters (Hu et al., 2022; Han et al., 2023).

### 3.6 Training Objective

Training optimizes standard causal language modeling over:

$$\texttt{prompt} \,\|\, \texttt{answer} \,\|\, \langle \text{EOS} \rangle,$$

with loss applied only to answer tokens. This setup aligns training and inference formats across all model variants.

### 3.7 Prompting Baselines

Zero-shot and ICL baselines use the same prompt template and decoding procedure as supervised models, but without parameter updates. All methods share the same prediction target, normalization procedure, and evaluation metric to ensure direct comparability.

Random ICL uses a fixed sampled demonstration set, while BM25 ICL retrieves demonstrations dynamically for each test instance.

**Decoding.** Greedy decoding is used with `max_new_tokens=8`, which is sufficient to generate a valid ICD prefix.

**Cross-dataset transfer.** To evaluate robustness under distribution shift, models trained on MIMIC-IV are evaluated directly on the external eICU dataset (Pollard et al., 2018) without target-domain adaptation. The same prediction and normalization protocols are used to ensure fair comparison across datasets.

## 4 Experiments

**Dataset.** Experiments are conducted on a leakage-reduced admission-time diagnosis benchmark derived from MIMIC-IV (Johnson et al., 2023). Each instance consists of an admission-time clinical note paired with a three-character ICD prefix label, with the label space restricted to the 50 most frequent prefixes identified prior to patient-level splitting ($|\mathcal{Y}| = 50$). Data are split by subject ID into train, validation, and test sets to prevent patient overlap, with one prediction instance per admission. Dataset statistics are provided in Appendix Table 7. All main results are reported on the held-out test split ($n = 13{,}516$).

For cross-dataset evaluation, an external test set is constructed from the *eICU Collaborative Research Database* (Pollard et al., 2018). Clinical records from the first 24 hours of each ICU stay are processed under the same task formulation and label space as MIMIC-IV. Models trained on MIMIC-IV are evaluated directly on eICU without further fine-tuning or target-domain adaptation.

**Evaluation Metric.** Performance is measured using exact-match (EM) accuracy over normalized ICD prefixes under greedy decoding. All main results use `max_new_tokens=8`. Supplementary strict-versus-relaxed analyses are provided in Appendix Table 12 to verify that performance gaps are not driven by output-formatting artifacts.

**Models.** Qwen2.5-7B-Instruct (Team, 2024) serves as the primary backbone for the main comparisons. Experiments evaluate zero-shot prompting, in-context learning (ICL), and supervised fine-tuning with LoRA (Hu et al., 2022).

LoRA is additionally evaluated on Meta-Llama-3-8B-Instruct (AI@Meta, 2024), Gemma-1.1-7B-IT, and Gemma-2-9B-IT (Team and DeepMind, 2024; Team et al., 2024). Retrieval and encoder baselines include BM25+kNN (Robertson and Zaragoza, 2009), dense kNN with frozen BioClinicalBERT embeddings (Alsentzer et al., 2019), and a supervised BioClinicalBERT classifier.

**Implementation Details.** LoRA is applied with rank $r = 16$, $\alpha = 32$, and dropout $0.05$. Models are trained using the AdamW optimizer with a learning rate of $2 \times 10^{-5}$ for 2 epochs and an effective batch size of 16. The maximum sequence length is set to 1024 tokens, with left truncation applied when necessary.

**Inference.** Greedy decoding is used for all main experiments with `max_new_tokens=8`, which is sufficient for ICD-prefix generation. Beam-search analyses (e.g., top-3 accuracy) are reported separately in Appendix Table 15 and do not affect the primary EM metric.

## 5 Results

### 5.1 Main Results

Table 1 summarizes the primary diagnosis results. Unless otherwise specified, all comparisons use the same test split, greedy decoding protocol, and ICD-prefix normalization rule.

**Prompting vs. Supervised Adaptation.** Under fixed prompting protocols, LLMs demonstrate a clear performance ceiling: zero-shot prompting yields only 2.45% EM, and 1-shot ICL reaches at most 12.61% even with structured inputs. In contrast, supervised adaptation via LoRA substantially improves performance, reaching 67.51% EM with structured admission-time representations.

Crucially, our error analysis (Appendix Table 12) shows that strict and relaxed exact match metrics are identical for both zero-shot and 1-shot prompting. This indicates that the poor performance of prompting is not caused by output-formatting or normalization artifacts, but instead reflects the difficulty of performing diagnosis prediction through prompting alone.

**Impact of Input Representation.** Input organization substantially influences diagnostic accuracy. Moving from default truncated notes to admission-only information provides a modest improvement for LoRA (56.21% to 58.40%). The largest gain occurs when adopting structured admission-time representations, which improve LoRA performance to 67.51% (+9.11% over raw admission text). These results suggest that the organization of clinical evidence plays a substantial role in effective model adaptation. The following analyses examine the generality of these trends across backbones and compare them against retrieval and encoder-based baselines.

**Generality Across Backbones.** To verify that the observed trends are not specific to a single backbone, the structured admission-time setup is evaluated across multiple model families (Table 2). Meta-Llama-3-8B-Instruct achieves 72.33% EM, while Gemma-2-9B-IT reaches the highest perfor-

| Method | Input Representation | Backbone | EM (%) |
|---|---|---|---|
| Zero-shot | Default (Truncated) | Qwen2.5-7B-Instruct | 2.45 |
| 1-shot ICL | Default (Truncated) | Qwen2.5-7B-Instruct | 11.85 |
| 1-shot ICL | Admission-only | Qwen2.5-7B-Instruct | 11.73 |
| 1-shot ICL | Structured Admission-time | Qwen2.5-7B-Instruct | 12.61 |
| LoRA | Default (Truncated) | Qwen2.5-7B-Instruct | 56.21 |
| LoRA | Admission-only | Qwen2.5-7B-Instruct | 58.40 |
| LoRA | Structured Admission-time | Qwen2.5-7B-Instruct | 67.51 |

Table 1: Main diagnosis results on the MIMIC-IV test set ($n$ = 13,516). We compare default truncated notes, admission-only text, and structured admission-time representations across prompting and supervised adaptation settings.

| Backbone | EM (%) |
|---|---|
| Qwen2.5-7B-Instruct (Team, 2024) | 67.51 |
| Meta-Llama-3-8B-Instruct (AI@Meta, 2024) | 72.33 |
| Gemma-1.1-7B-IT (Team and DeepMind, 2024) | 71.13 |
| Gemma-2-9B-IT (Team et al., 2024) | **75.44** |

Table 2: LoRA results across instruction-tuned backbones using the structured admission-time setup. The consistent gains across model families suggest that the observed trends are not backbone-specific.

| Method | EM (%) |
|---|---|
| BM25+kNN | 45.26 |
| Dense kNN (frozen BioClinicalBERT) | 39.26 |
| Encoder + linear head | 67.34 |
| Qwen LoRA | **67.51** |

Table 3: Retrieval and encoder-based baselines on the structured admission-time test set. Dense kNN reports the best frozen BioClinicalBERT retrieval setting among the sweeps ($k$=50).

mance at 75.44% EM. The consistent improvements across architectures indicate that the benefits of structured admission-time representations and supervised adaptation generalize across instruction-tuned LLM families.

Further experiments with 3-shot ICL under longer-context evaluation (Appendix Table 17) do not yield substantial improvements, suggesting that increasing prompt complexity alone is insufficient to bridge the gap between prompting and supervised adaptation in this setting.

### 5.2 Retrieval and Encoder Baselines

Table 3 compares retrieval and encoder-based baselines on the structured admission-time test set. BM25+kNN reaches 45.26% EM and dense kNN 39.26%, indicating the presence of a substantial case-similarity signal in admission-time diagnosis.

A supervised BioClinicalBERT encoder with mean pooling and a linear classification head achieves 67.34% EM, approaching the Qwen LoRA result (67.51%).

This suggests that supervised encoder models already capture substantial diagnostic signal under the current formulation, where prediction is constrained to a fixed set of coarse ICD prefixes.

The relatively strong BM25 performance also supports the use of retrieval-based demonstration selection in later ICL experiments.

### 5.3 Information Source Analysis

Table 4 analyzes how different clinical information sources contribute to diagnosis under two complementary protocols: evaluation-time ablation and matched train+test ablation.

Under evaluation-time ablation, HPI alone achieves 56.59% EM, showing that narrative clinical history provides the dominant diagnostic signal. In contrast, Pertinent Results alone reaches only 20.86% EM. Using CC+HPI+PMH yields 59.71%, while adding Pertinent Results increases performance to 67.40%, close to the full structured model (67.51%). Removing Physical Examination causes almost no degradation. This may partly reflect the highly templated or predominantly normal nature of physical examination documentation in many admission notes.

Matched train+test supervision reveals a similar pattern. HPI-only training reaches 61.59%, while Results-only improves to 41.19%, indicating that objective findings become more useful when the model is optimized specifically for them, although they remain substantially weaker than narrative history.

Overall, the results suggest a clear hierarchy of information usage: narrative clinical history (especially HPI) provides the dominant signal, PMH offers modest gains, and objective findings serve pri-

| Section setting | Eval-only EM (%) | Matched train+test EM (%) |
|---|---|---|
| HPI only | 56.59 | 61.59 |
| Results only | 20.86 | 41.19 |
| CC + HPI + PMH | 59.71 | 63.05 |
| CC + HPI + PMH + Results | 67.40 | 66.88 |
| Full structured representation | **67.51** | — |

Table 4: Section ablations on the structured admission-time representation. Eval-only ablation uses a model trained on the full structured input and perturbs only the test-time input. Matched train+test trains and evaluates each model on the same ablated input.

| Input design | EM (%) |
|---|---|
| Raw truncation | 64.14 |
| Structured, no compression | **69.06** |
| Structured, reverse order | 66.33 |
| Structured, shuffled order | 67.37 |
| Full structured v2 | 67.51 |

Table 5: Input design ablations with retraining. Section-based structure consistently outperforms raw truncation. These results suggest that explicit organization contributes more to performance than heuristic filtering.

| Approach | EM | Same Block |
|---|---|---|
| Zero-shot | 0.11 | 1.32 |
| LoRA transfer | 35.30 | 49.93 |

Table 6: Cross-dataset transfer on eICU. Same Block denotes predictions falling within the same coarse ICD block as the gold label.

marily as complementary refinement, while Physical Examination contributes minimally under the current formulation.

### 5.4 Input Design Analysis

Table 5 isolates which aspects of the input design contribute most to performance. Compared to raw truncated text, all structured variants improve exact-match accuracy: raw truncation reaches 64.14% EM, while structured variants range from 66.33% to 69.06%.

Among the Qwen2.5-7B-Instruct variants, the no-compression setting performs best at 69.06%, outperforming the full structured representation (67.51%). These results suggest that explicit organization of clinical evidence contributes more to performance than heuristic compression of objective findings. Notably, the no-compression variant is both longer and more heavily truncated than the full structured representation, yet still achieves higher EM. Under the same 1024-token limit used during training, 83.3% of no-compression examples are truncated at test time, compared with 67.3% for the full structured representation (Appendix Table 13). This suggests that structured organization may help preserve diagnostically salient information even under substantial truncation.

To examine the role of ordering, the default clinically motivated section order is compared with alternative section orderings. Reversing the section order yields 66.33%, while shuffling sections gives 67.37%. The relatively small differences suggest that preserving semantically separated sections is more important than any single fixed ordering.

A supplementary leakage-masking analysis is provided in Appendix Table 14. Rule-based masking of residual diagnosis-like cues reduces EM only slightly from 67.51% to 66.71%, suggesting that simple lexical shortcuts are not the primary driver of performance.

### 5.5 Cross-Dataset Evaluation on eICU

To evaluate generalization beyond a single institution, models are tested under cross-dataset transfer on eICU. As shown in Table 6, zero-shot inference without task adaptation performs poorly, achieving only 0.11% EM and 1.32% coarse-block accuracy.

In contrast, the MIMIC-trained LoRA model reaches 35.30% EM under direct transfer, with Same Block accuracy increasing to 49.93%, indicating partial retention of coarse diagnostic consistency despite substantial domain shift.

Compared with the in-domain results (67–75% EM), performance drops substantially under cross-dataset evaluation. Retrieval-based methods degrade even more sharply: BM25+kNN falls from 45.26% to 23.82% EM, suggesting that retrieval relies more heavily on dataset-specific lexical overlap, whereas supervised models learn more transferable diagnostic representations.

Error analysis further shows that some cross-dataset mismatches arise from semantically aligned ICD-9/ICD-10 mappings, suggesting that strict

exact-match evaluation may provide a lower bound on transfer performance across differing coding systems.

### 5.6 Error Analysis

We analyze errors on the main Qwen2.5-7B-Instruct LoRA model under the structured admission-time setting (Appendix Table 12). Strict and relaxed exact match are identical at 67.51%, showing that the reported EM is not driven by output-formatting artifacts. Generation failures are negligible (0.03%), indicating that most remaining errors arise from diagnosis prediction rather than generation failure. Using a coarse ICD block mapping, same-block accuracy reaches 74.22%, showing that many remaining errors fall within clinically related diagnostic categories. Among all test examples, 6.70% fall within the correct coarse ICD block but predict the wrong prefix, whereas 25.75% fall into a different coarse block. Common confusions are concentrated among clinically related diagnoses rather than uniformly distributed. For example, heart failure (ICD-9: 428) is frequently confused with myocardial infarction (ICD-9: 410). Similarly, heart failure (ICD-10: I50) is often predicted as hypertensive heart disease (ICD-10: I11/I13), reflecting the close clinical relationship between these conditions. These patterns suggest that the model captures coarse clinical structure but still struggles with fine-grained distinctions among closely related conditions.

## 6 Discussion

The results consistently suggest that reliable clinical diagnosis with LLMs depends more on supervised learning and information organization than on prompting alone. Across model families, prompting-based approaches remain substantially weaker than supervised adaptation, and additional analyses indicate that this gap is not driven by output-formatting artifacts.

Input representation also plays a central role. Structuring records into clinically meaningful sections consistently improves performance over raw truncated text. Further ablations suggest that these gains arise primarily from improved structure rather than aggressive filtering, as heuristic compression of objective findings is not consistently beneficial.

A clear hierarchy of information usage emerges. Narrative clinical history, particularly the HPI, carries most of the diagnostic signal, while laboratory and imaging findings provide complementary refinement. Physical examination contributes comparatively little in this setting. This pattern resembles clinical diagnostic practice, where narrative history establishes initial hypotheses and objective findings are used to refine nearby distinctions.

The comparison between retrieval and supervised models further suggests that diagnosis contains both strong case-similarity signal and transferable supervised structure. BM25 achieves competitive performance, indicating that many cases can be partially resolved through lexical similarity alone, yet retrieval degrades more sharply under cross-dataset transfer. In contrast, supervised representations exhibit stronger robustness across datasets, suggesting that supervised adaptation captures more transferable diagnostic structure beyond surface-level lexical matching.

Finally, model errors are highly structured rather than random. Misclassifications tend to occur among clinically related diagnoses, suggesting that models often capture coarse disease structure while struggling with fine-grained distinctions between nearby conditions.

## 7 Conclusion

This work studies clinical diagnosis with LLMs under a controlled admission-time setting. Prompting-based approaches perform poorly, while supervised adaptation substantially improves performance across model families. Structured inputs consistently outperform raw text, highlighting the importance of organizing clinical evidence. Analyses further reveal a hierarchy of information usage: narrative history dominates diagnosis, while objective findings provide complementary refinement. Effective diagnosis from clinical notes depends not only on model capacity, but also on supervision and input structure.

## Limitations

This study has several limitations. It focuses on coarse ICD prefix prediction, which simplifies the diagnostic task and does not fully capture fine-grained or multi-label clinical complexity, although coarse prefixes provide a more stable setting for controlled analysis. In addition, diagnosis labels are derived from retrospective clinical coding processes, which may not perfectly reflect real-time diagnostic reasoning at admission. Models are primarily trained on MIMIC-IV, and generalization to other institutions and populations remains limited despite cross-dataset evaluation. The structured input pipeline relies on heuristic section extraction and result processing, which may omit relevant information or introduce bias. Finally, the study does not evaluate real-world deployment or clinician interaction.

## Ethics Statement

This work uses de-identified clinical data from the MIMIC-IV dataset and is intended solely for research purposes. We do not study deployment in real clinical environments, and the models evaluated in this paper are not suitable for direct medical decision-making.

There are several ethical risks in applying large language models to clinical diagnosis. Incorrect predictions may lead to misdiagnosis, delayed treatment, or over-reliance on automated systems. In addition, model behavior may not generalize across institutions, populations, or documentation styles, which could exacerbate disparities if such systems were used without careful validation.

Accordingly, we emphasize that these models should not be used for clinical decision-making without rigorous external validation, prospective evaluation, and oversight by qualified healthcare professionals. Our goal is to better understand how LLMs use clinical information under controlled conditions, rather than to advocate immediate deployment in practice.

## References


AI@Meta. 2024. Llama 3 model card.

Kitti Akkhawatthanakun, Lalita Narupiyakul, Konlakorn Wongpatikaseree, Narit Hnoohom, Chakkrit Termritthikun, and Paisarn Muneesawang. 2026. Integrating agentic artificial intelligence to automate international classification of diseases, tenth revision, medical coding. In *Informatics*, volume 13, page 39. MDPI.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*, pages 72–78.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hang Dong, Víctor Suárez-Paniagua, William Whiteley, and Honghan Wu. 2021. Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation. *Journal of biomedical informatics*, 116:103728.

Raffaello Fornasiere, Nicolò Brunello, Vincenzo Scotti, and Mark Carman. 2024. Medical information extraction with large language models. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 456–466.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexei Figueroa, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

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

Kundan Krishna, Sopan Khosla, Jeffrey P Bigham, and Zachary C Lipton. 2021. Generating soap notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972.

Eric Lehman. 2024. *Practical Considerations For the Deployment of Clinical NLP Systems*. Ph.D. thesis, Massachusetts Institute of Technology.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns*, 5(3).

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. 2018. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):180178.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*, volume 4. Now Publishers Inc.

Tiago Rodrigues and Carla Teixeira Lopes. 2025. Large language model-based generation of discharge summaries. *arXiv preprint arXiv:2512.06812*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Gemma Team and Google DeepMind. 2024. Gemma: Open models based on gemini research and technology.

Gemma Team, Morgane Riviere, Shreya Pathak, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

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

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. 2024. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138.

Bharath Velamala, Elham Sagheb Hossein Pour, Michael Lin, and Jungwei Wilfred Fan. 2025. medspacyv: a graphical user interface for the open source medspacy natural language processing package. *JAMIA open*, 8(4):ooaf094.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Weipeng Zhou, Majid Afshar, Dmitriy Dligach, Yanjun Gao, and Timothy Miller. 2023. Improving the transferability of clinical note section classification models with bert and large language model ensembles. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 125–130.

| Split | Admissions ($N$) |
|---|---|
| Train | 106,279 |
| Validation | 13,439 |
| Test | 13,516 |
| **Total** | **133,234** |

Table 7: Dataset statistics. Each admission corresponds to one diagnosis instance. The label space is fixed to 50 ICD prefixes across all splits.

# A Dataset Construction and Preprocessing

## A.1 Data Split

Dataset split statistics are summarized in Table 7.

## A.2 Section Anchors

Table 8 summarizes the header anchors used for admission-time input construction.

Anchors are implemented as case-insensitive regular expressions. Headers that do not match predefined anchors are not treated as section boundaries and therefore remain inside the nearest matched section, which is a limitation of the deterministic header-based pipeline.

## A.3 Leakage Control Pipeline

See Table 9 for the leakage-control procedures used during dataset construction.

Leakage control combines structural section removal with supplementary lexical masking checks.

## A.4 Segmentation Fallback Statistics

Table 10 summarizes segmentation fallback statistics for the header-driven input pipeline.

Segmentation fallback is extremely rare across all splits. All fallback cases map to `admission_snapshot`; no cases fall back only to `initial_info`.

## A.5 Pertinent Results Compression Fallback

Table 11 summarizes fallback behavior in the Pertinent Results compression pipeline.

This fallback is distinct from segmentation fallback and is intended to preserve information when heuristic filtering removes excessive content.

## A.6 Truncation Profile Under the 1024-Token Limit

Table 13 summarizes prompt-length statistics under the 1024-token limit used during LoRA training.

Structured admission variants are frequently truncated under the 1024-token limit. The no-compression setting is both longer and more heavily truncated than the default structured input, yet still achieves higher exact-match accuracy.

## A.7 Leakage Check

Table 14 summarizes the supplementary leakage-masking analysis.

Rule-based masking targets residual diagnosis-like lexical shortcuts, including impression-style phrases, ICD-like patterns, and templated diagnosis language. Exact-match accuracy decreases only slightly after masking.

# B Additional Evaluation Analyses

## B.1 Top-3 Diagnostic Accuracy

Table 15 reports top-3 diagnostic accuracy under beam decoding.

Beam decoding substantially improves coarse diagnostic recall compared with greedy exact match.

## B.2 Error Decomposition Across Prompting and LoRA

Table 12 summarizes error decomposition across prompting and LoRA settings on the structured admission v2 test set.

Strict and relaxed EM are identical across prompting and LoRA settings, indicating that output-formatting artifacts are negligible under the current evaluation protocol.

## B.3 BM25 versus LoRA on the Test Set

Table 16 summarizes test-set agreement between BM25+kNN and the main LoRA model.

LoRA correctly predicts 4011 cases missed by BM25, while BM25 correctly predicts 1003 cases missed by LoRA, indicating partially complementary behavior between retrieval and supervised models.

## B.4 Three-Shot Prompting Under a Long Context Window

Table 17 reports prompt-only performance under longer-context evaluation.

Even after removing the earlier truncation artifact, 3-shot ICL remains below 1-shot prompting under long-context evaluation.

| Role | Internal section | Header anchors |
|---|---|---|
| Keep | chief_complaint | Chief Complaint |
| Keep | hpi | History of Present Illness |
| Keep | pmh | Past Medical History |
| Keep | pertinent_results | Pertinent Results; Significant Results |
| Keep | physical_exam | Physical Exam; Physical Examination |
| Drop | course | Brief Hospital Course; Short Hospital Course; Hospital Course; ICU Course |
| Drop | discharge_diagnosis | Discharge Diagnosis; Discharge Diagnoses |
| Drop | discharge_medications | Medications on Discharge; Discharge Medications |
| Drop | surgical | Major Surgical or Invasive Procedure; Past Surgical History |
| Drop | studies | Studies on This Admission |
| Drop | interval | Interval History |
| Drop | metadata | Service; Allergies; Attending |

Table 8: Header anchors used for admission-time input construction.

| Step | Input object | Operation | Usage |
|---|---|---|---|
| Section removal | Raw discharge summary | Remove retrospective sections | Main dataset construction |
| Lexical masking | Admission-time input | Mask diagnosis-like phrases | Robustness check |

Table 9: Leakage-control procedures used during dataset construction.

| Split | Fallback count | Rate (%) |
|---|---|---|
| Train | 27 | 0.0254 |
| Validation | 6 | 0.0447 |
| Test | 5 | 0.0370 |

Table 10: Segmentation fallback statistics for the header-driven input pipeline. When no valid structured span is recovered, the note falls back to `admission_snapshot`.

| Split | Fallback (%) | Trunc-hit (%) |
|---|---|---|
| Train | 4.7 | 0.05 |
| Validation | 4.7 | 0.05 |
| Test | 4.7 | 0.05 |

Table 11: Fallback behavior in the Pertinent Results compression pipeline. "Fallback" denotes cases where heuristic filtering removed excessive content and the original result span was restored. "Trunc-hit" denotes fallback cases that still exceeded the 1024-token limit after restoration.

# C Cross-Dataset Supplementary Results

## C.1 Prompt-Based Cross-Dataset Evaluation on eICU

Table 18 reports prompt-based cross-dataset evaluation on eICU.

Prompt-based evaluation on eICU is performed on sampled subsets because of the computational overhead of large-scale long-context prompting.

# D Prompt Templates and In-Context Setup

The following templates are used for prompt-based experiments.

```
### Instruction:
You are a clinical reasoning assistant. Predict
    only the coarse ICD prefix
(3-character code) for the primary diagnosis
    from the partial discharge summary.
Return only the prefix, with no extra
    explanation.

### Input:
{partial_note}

### Answer:
```

**In-Context Template** Each demonstration block has the following form:

```
### Input:
{demo_note_i}

### Answer:
{PREFIX_i}
```

The final test instance is appended as:

```
### Input:
{test_note}

### Answer:
```

**Demonstration Selection** For random ICL, a fixed set of demonstrations is sampled once from the training split and reused for all test examples.

| Setting | Strict EM (%) | Relaxed EM (%) | Same-block (%) | Same-block wrong prefix (%) | Different block (%) | No ICD-like token (%) |
|---|---|---|---|---|---|---|
| Zero-shot | 2.29 | 2.29 | 7.86 | 5.57 | 91.68 | 0.44 |
| 1-shot ICL | 12.61 | 12.61 | 36.00 | 23.39 | 63.58 | 0.42 |
| LoRA | 67.51 | 67.51 | 74.22 | 6.70 | 25.75 | 0.03 |

Table 12: Error decomposition on the structured admission v2 test set. "Same-block wrong prefix" denotes predictions that fall within the correct coarse ICD block but fail to match the exact ICD prefix.

| Input variant | Mean prompt tokens | Median | P95 | Truncated (%) | Mean kept ratio |
|---|---|---|---|---|---|
| Default | 207 | 198 | 279 | 0.13 | ≈1.00 |
| Admission | 274 | 265 | 375 | ≈0.01 | ≈1.00 |
| Structured admission v2 | 1312 | 1267 | 2302 | 67.3 | 0.79 |
| Structured admission v2, no compression | 1845 | 1742 | 3513 | 83.3 | 0.63 |
| Raw truncation | 993 | 967 | 1693 | 44.4 | 0.90 |

Table 13: Prompt-length profile under the same tokenizer and preprocessing used for LoRA training with maximum sequence length 1024.

| Setting | EM (%) |
|---|---|
| Full structured v2 | 67.51 |
| Leakage-masked input | 66.71 |

Table 14: Leakage check using rule-based masking on top of the already leakage-reduced structured admission input.

| Metric | Score (%) |
|---|---|
| Exact Match (greedy) | 67.51 |
| Top-3 Accuracy (beam) | 88.70 |

Table 15: Top-3 diagnostic accuracy under beam decoding.

For BM25 ICL, demonstrations are retrieved separately for each test example using the same tokenization as the BM25 retrieval baseline.

| | LoRA miss | LoRA hit |
|---|---|---|
| BM25 miss | 3388 | 4011 |
| BM25 hit | 1003 | 5114 |

Table 16: Test-set agreement between BM25+kNN and the main LoRA model on structured admission v2.

| Setting | EM (%) |
|---|---|
| 1-shot ICL | 12.61 |
| 3-shot ICL (32768 context) | 6.50 |

Table 17: Prompt-only performance on structured admission v2 under longer-context evaluation.

| Evaluation subset | Zero-shot | 1-shot | 3-shot | LoRA |
|---|---|---|---|---|
| 5k samples | 0.12 | 15.32 | 24.64 | 35.84 |
| 10k samples | 0.11 | 15.44 | 24.09 | 35.48 |

Table 18: Prompt-based cross-dataset evaluation on eICU. Results are reported as exact-match accuracy (%).