MAX-EVAL-11: A Large-Scale Benchmark for Evaluating Large Language Models on Full-Spectrum ICD-11 Medical Coding

Ujjwal Singh
Max Healthcare
ujjwal.singh@maxhealthcare.com

Sarthak Deshwal
Plaid Inc.
sarthakdeshwal@duck.com

Nitish Dube
Max Healthcare
ntish.dube@maxhealthcare.com

Arjun Sharma
Max Healthcare
arjun.sharma@maxhealthcare.com

Abstract

The global transition to the ICD-11 taxonomy demands robust automated medical coding, yet comprehensive benchmarks to evaluate Large Language Models (LLMs) on this task remain absent. We introduce MAX-EVAL-11 [1] as a foundational resource, the first large-scale benchmark for full-spectrum ICD-11 medical coding. MAX-EVAL-11 comprises 10,000 MIMIC-III discharge summaries with mapped, expert-validated subset (12%) ICD-11 annotations spanning 99.87% of the diagnostic taxonomy. To better reflect clinical utility, we propose a novel hierarchical evaluation framework that assigns partial credit based on ICD-11's 5-level structure, addressing the brittleness of traditional exact-match metrics.

To illustrate benchmark utility, we evaluate LLMs on a stratified 500-patient sample, revealing significant performance gaps. The best-performing model (Claude 4 Sonnet) achieves a weighted score of 0.433, outperforming both general-purpose peers and specialized medical models (MedCoder). Crucially, all models exhibit near-zero exact match rates (0–4.8%) and rely primarily on hierarchical credit, underscoring the extreme difficulty of precise ICD-11 code generation. Furthermore, the superiority of general-purpose LLMs over legacy ICD-10 medical models (with ICD-11 codelist) suggests that broad reasoning capabilities currently outweigh domain-specific training for complex taxonomy scaling.

1 Introduction

Medical coding, transforming clinical narratives into standardized diagnostic codes is fundamental to modern healthcare for statistical analysis, reimbursement, and epidemiological research. The International Classification of Diseases (ICD), maintained by the WHO, provides the global standard for this task.

ICD-11, officially adopted in 2019 and effective since January 2022, represents a major evolution in medical classification (World Health Organization, 2019). As of 2024, 132 member states are implementing ICD-11, which introduces over 55,000 unique codes with post-coordination for granular clinical documentation (Harrison et al., 2021); (World Health Organization, 2024). Manual medical coding faces substantial challenges from growing documentation volumes, evolving regulations, and workforce shortages. The ICD-11 transition intensifies these pressures, requiring complex crosswalk mapping. Only 23.5% of ICD-9 codes achieve direct one-to-one ICD-11 mapping, complicating automated transition (Fung et al., 2021).

Recent AI advances, particularly large language models (LLMs), have sparked interest in automated medical coding. However, specialized models achieve F1 scores of only 0.539 to 0.719 on ICD-10 benchmarks (Baksi et al., 2025; Edin et al., 2023), while general-purpose LLMs face additional challenges (Hou et al., 2025). ICD-11's multidimensional taxonomy—capable of generating vast combinations via post-coordination—presents even greater complexity. Despite progress, existing benchmarks primarily focus on ICD-9 and ICD-10. MIMIC-III (Johnson et al., 2016) and MIMIC-IV-ICD (Nguyen et al., 2023) provide robust evaluation for legacy systems, but no comprehensive ICD-11 benchmark exists. Furthermore, traditional precision, recall, and F1 metrics inadequately capture the hierarchy of clinical relevance: accurate identification of a primary diagnosis is more critical than secondary conditions, yet existing metrics weight all predictions equally (Mullenbach et al., 2018).

We present **MAX-EVAL-11**, a comprehensive benchmark for evaluating LLM performance on ICD-11 medical coding. Our contributions are: **First**, a large-scale dataset with 10,000 real MIMIC-III discharge summaries with systematic

[1] Dataset: https://huggingface.co/datasets/mas-namtla/max-eval-11 (gated access).

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 282–291

ICD-9 to ICD-11 conversion and complete taxonomy coverage. **Second**, a novel rank-weighted precision scoring methodology that assigns differential importance based on clinical relevance and diagnostic hierarchy, better reflecting real-world accuracy requirements. **Third**, baseline performance metrics using state-of-the-art LLMs (Claude, Gemini), demonstrating significant improvement opportunities in automated ICD-11 coding.

2 Related Work

2.1 Deep Learning Approaches for ICD Coding

Early neural approaches (RNNs, CNNs) established feasibility of automated coding but struggled with extreme multi-label classification (13,000+ codes) and long clinical documents (Choi et al., 2017; Baumel et al., 2018). Mullenbach et al. (2018) introduced CAML, using per-label attention to identify diagnosis-relevant text segments, establishing the label-wise attention paradigm for ICD coding. Subsequent refinements include LAAT (Vu et al., 2020), pseudo label-wise attention (Wu et al., 2021), and HiLAT (Liu et al., 2022), progressively improving efficiency and hierarchical coding accuracy. The integration of pretrained language models brought further improvements. Huang et al. (2022) introduced PLM-ICD, incorporating domain-specific pretraining, segment pooling, and specialized label attention, achieving 59.8% micro-F1 on MIMIC-III. Liu et al. (2023) extended these capabilities with extreme multi-label transformers adapted for long clinical narratives (ClinicalBIGBIRD), reaching 60.8% micro-F1. However, these advances remain constrained to ICD-9 and ICD-10 evaluation, with no comprehensive ICD-11 benchmarks available despite the classification's 2019 WHO adoption and ongoing global implementation across 132 member states. Wu et al. (2025) introduced MIMIC-SR-ICD11, a dataset derived from MIMIC-IV for narrative-based ICD-11 diagnosis, though limited to primary diagnosis only and covering a narrow subset of the full taxonomy without a hierarchical evaluation framework.

2.2 Evaluation Frameworks and Limitations

Kim et al. (2022) introduced AnEMIC, providing standardized preprocessing and evaluation protocols for MIMIC-III benchmarks—an important step toward reproducibility. However, three fundamental limitations persist across existing evaluation frameworks, directly motivating MAX-EVAL-11's development.

Metric mismatch with clinical hierarchy. Standard precision, recall, and F1 metrics treat all diagnostic codes uniformly despite differing clinical salience (Kim et al., 2022). This approach fails to capture medical classification systems' hierarchical nature. For instance, predicting "Infectious Disease" (ICD-11 chapter 1A00-1G9Z) when the ground truth is "Bacterial Pneumonia" (CA40.0) receives zero F1 credit, yet the model correctly identified the broad disease category (infection) and affected organ system (respiratory)—partial accuracy valuable in clinical practice where physicians can narrow diagnoses from correct disease families. Moreover, the distinction between primary and secondary diagnoses carries critical clinical importance that flat metrics ignore: primary diagnoses drive treatment decisions and resource allocation, making errors in primary identification far more consequential than missing secondary comorbidities (e.g., failing to identify acute myocardial infarction as primary could delay life-saving interventions). Existing metrics weight all predictions equally, obscuring these safety-critical distinctions.

The benchmark-practice performance gap. Recent evaluations reveal troubling disconnects between constrained benchmark performance and real-world capability. Hou et al. (2025) demonstrate that models achieving 97–99% exact match on simplified ICD-9 test sets collapse to 3.85–10.90% when evaluated on realistic clinical scenarios. This dramatic degradation stems from two factors: *extreme label spaces* where models must identify 10–15 relevant codes from 50,000+ candidates per clinical note (fundamentally different from constrained benchmarks limiting evaluation to hundreds of common codes), and *realistic multi-diagnosis complexity* in discharge summaries containing multiple diagnoses spanning organ systems, temporal progression (admission → complications → discharge), and complex interactions (e.g., diabetes exacerbating pneumonia recovery) that simplified benchmarks lack.

Absent ICD-11 coverage. Despite ICD-11's official WHO adoption in 2019 and implementation across 132 member states as of 2024 (World Health Organization, 2019), existing benchmarks remain focused on legacy ICD-9 (Johnson et al., 2016) and ICD-10 (Nguyen et al., 2023) classifications. The few ICD-11 pilot studies remain narrow in scope, typically evaluating models on small

taxonomy subsets (500–1,000 codes) or synthetic data lacking real clinical documentation complexity. No comprehensive benchmark spans ICD-11's full 55,000-code diagnostic taxonomy with authentic clinical text, creating critical evaluation gaps as healthcare systems worldwide transition to the new standard.

2.3 Positioning MAX-EVAL-11

MAX-EVAL-11 addresses these gaps through: (1) full-spectrum ICD-11 coverage across 10,000 MIMIC-III discharge summaries; (2) hierarchical evaluation metrics reflecting clinical diagnostic priority; and (3) comprehensive LLM baselines on realistic multi-diagnosis scenarios.

3 Methodology

3.1 ICD-9 to ICD-11 Mapping Framework

The transition from ICD-9 (≈13,000 codes) to ICD-11 (≈55,000 codes) represents a fundamental shift in medical classification granularity (World Health Organization, 2019). Given an ICD-9 code $c_9 \in \mathcal{C}_9$, we seek to identify the optimal ICD-11 code set $S_{11} \subseteq \mathcal{C}_{11}$ that preserves clinical semantics. Our hybrid approach combines semantic embedding-based candidate retrieval (efficient search space reduction from 55,000+ codes) with LLM-based contextual selection (clinical disambiguation of related codes).

3.2 Hybrid Semantic-LLM Mapping Pipeline

Problem Formulation. Let $\mathcal{D}_9 = \{(c_9^{(i)}, t_9^{(i)})\}_{i=1}^{N_9}$ and $\mathcal{D}_{11} = \{(c_{11}^{(j)}, t_{11}^{(j)})\}_{j=1}^{N_{11}}$ represent the ICD-9 and ICD-11 code sets, where c denotes the code identifier, t denotes the textual description, $N_9 \approx 13{,}000$, and $N_{11} \approx 55{,}000$. Our mapping function is:

$$f : \mathcal{C}_9 \to 2^{\mathcal{C}_{11}} \times [1, 10] \tag{1}$$

where $f(c_9) = (S_{11}, \sigma)$ returns ICD-11 codes S_{11} with LLM-assigned quality score $\sigma \in [1, 10]$.

Semantic Candidate Generation. While WHO provides partial ICD-9–ICD-11 crosswalk tables, these cover only 23.5% of ICD-9 codes with direct mappings (Fung et al., 2021); our hybrid pipeline addresses the remaining 76.5% for which no official mapping exists. We employ Bio_ClinicalBERT (Alsentzer et al., 2019), pretrained on 2 million MIMIC-III clinical notes, achieving F1 = 0.89 on i2b2 2010 medical entity recognition. Its 768-dimensional embeddings capture fine-grained semantic distinctions essential for differentiating clinically related diagnostic codes. We use the pretrained model without fine-tuning, as the task requires general medical semantic similarity rather than task-specific classification.

For each ICD-9 code, we construct composite inputs via string concatenation ($\oplus$):

$$x_9^{(i)} = c_9^{(i)} \oplus \text{“ - ”} \oplus t_9^{(i)} \tag{2}$$

and encode using Bio_ClinicalBERT's encoder ϕ : String $\to R^{768}$:

$$e_9^{(i)} = \phi(x_9^{(i)}) \tag{3}$$

Similarly, we encode all ICD-11 descriptions to obtain embeddings $e_{11}^{(j)}$. We identify top-$k = 10$ candidates via cosine similarity:

$$\text{sim}(e_9^{(i)}, e_{11}^{(j)}) = \frac{e_9^{(i)} \cdot e_{11}^{(j)}}{\|e_9^{(i)}\| \|e_{11}^{(j)}\|} \tag{4}$$

where $k = 10$ balances efficiency with coverage for one-to-many mappings common in ICD-9 to ICD-11 transitions.

LLM-Based Contextual Selection. Top-10 candidates often contain clinically related but distinct codes (e.g., bacterial vs. viral pneumonia) requiring medical reasoning to disambiguate. We employ Gemini 2.0 Flash, selected after comparing GPT-4, Claude 3.7, and Gemini 2.0 on 100 sample mappings: Gemini achieved 87% expert agreement versus 79% (GPT-4) and 82% (Claude 3.7). Its structured output capabilities and 1-million token context enable processing complete taxonomies. We formulate prompts:

$$\mathcal{P}(c_9, \mathcal{C}^{\text{cand}}) = [\text{Task}_{\text{instr}}, \text{Code}_{\text{ctx}}, \text{Output}_{\text{fmt}}] \tag{5}$$

where $\text{Task}_{\text{instr}}$ instructs the model to *rate the clinical relevance of each ICD-11 candidate for the given ICD-9 code on a scale of 1–10 (10 = exact semantic match; 1 = unrelated)*; Code_{ctx} includes ICD-9 code c_9 with the 10 ICD-11 candidates and their descriptions; and $\text{Output}_{\text{fmt}}$ specifies a JSON schema of `{code: string, score: int}` per candidate. The LLM returns $M_i = \{(c_j^{11}, \sigma_{ij})\}$ where $\sigma_{ij} \in [1, 10]$. We retain all candidates with $\sigma_{ij} \geq 7$ as the final mapping set S_{11}; if no candidate meets this threshold, the highest-scoring candidate is retained to ensure full coverage.

Methodological Independence. Label generation uses only ICD code taxonomies for code-to-code

translation (Gemini 2.0 Flash), while evaluation requires clinical narrative understanding from discharge summaries (Gemini 2.5 Flash), distinct models, tasks, and information sources (see Conclusion).

3.3 Quality Score-Based Stratification and Quality Assessment

We stratify by mapping quality score Σ_p, the average LLM-assigned score across patient p's mappings:

$$\Sigma_p = \frac{1}{|M_p|} \sum_{(c_9, c_{11}, \sigma) \in M_p} \sigma \tag{6}$$

where M_p represents all mappings for patient p. Binary stratification with $\tau = 7.0$ yields:

$$\text{Conf}(p) = \begin{cases} \text{High} & \text{if } \Sigma_p \geq 7.0 \\ \text{Low} & \text{if } \Sigma_p < 7.0 \end{cases} \tag{7}$$

This threshold was validated on 200 manual mappings: $\sigma \geq 7.0$ achieved 89% expert agreement versus 67% for $\sigma < 7.0$. The *High-Confidence Subset* ($\Sigma_p \geq 7.0$) optimizes for maximum reliability (billing, compliance), while *All-Matches Subset* enables robust evaluation across quality conditions.

We assess mapping quality via Coverage Rate $\rho = |\{p : |M_p| > 0\}|/N_p$ (proportion successfully mapped), Expansion Factor $\alpha = \sum_p |C_{11}^{(p)}| / \sum_p |C_9^{(p)}|$ (granularity increase, where $C_9^{(p)}$ and $C_{11}^{(p)}$ are code sets for patient p), and Semantic Preservation $\psi = \frac{1}{N_m} \sum_{i=1}^{N_m} \max_j \text{sim}(c_9^{(i)}, c_{11}^{(j)})$ (semantic alignment, where N_m is total mappings and $\text{sim}(\cdot, \cdot)$ is Equation 4). Values $\alpha > 1$ and $\psi > 0.75$ indicate successful ICD-11 expansion with semantic fidelity. The threshold $\psi > 0.75$ reflects the cosine similarity range at which BioClinicalBERT embeddings reliably preserve clinical semantics: values below 0.75 in our BioClinicalBERT embedding space correspond to pairs where the primary diagnostic category (e.g., infectious vs. cardiovascular) differs, as verified on the 200-mapping validation subset used for confidence threshold calibration (Section 3.3).

3.4 Hierarchical Evaluation Metrics

Traditional metrics assign binary credit (1 for exact match, 0 otherwise), ignoring ICD-11's hierarchical taxonomy. A model predicting "CA40.0Z Bacterial pneumonia, unspecified" for ground truth "CA40.0 Bacterial pneumonia" demonstrates clinical understanding despite technical mismatch. Similarly, primary diagnosis accuracy matters more than secondary coding, yet flat metrics weight equally. We propose multi-component scoring reflecting clinical priorities.

Hierarchical Credit Assignment. ICD-11's 5-level hierarchy (Chapter → Block → Category → Subcategory → Extension) enables partial credit $h(c_{\text{true}}, c_{\text{pred}})$ based on closest common ancestor:

$$h(c_{\text{true}}, c_{\text{pred}}) = \begin{cases} 1.0 & \text{exact} \\ 0.9 & \text{parent} \\ 0.8 & \text{grandparent} \\ 0.7 & \text{great-grandparent} \\ 0.6 & \text{chapter} \\ 0.0 & \text{unrelated} \end{cases} \tag{8}$$

Credits decrease uniformly by Δ=0.1 per taxonomy level from exact match (1.0) down to chapter-level match (0.6), reflecting ICD-11's 5-level hierarchy where each level represents a meaningful clinical distinction; codes sharing no chapter (e.g., an infectious disease predicted as a metabolic disorder) receive h=0.0 as they provide no clinically actionable signal regardless of proximity.

Metric Components. *Exact Match* (EM) measures precise coding:

$$\text{EM} = \frac{1}{N} \sum_{i=1}^{N} \frac{|P_i \cap G_i|}{|P_i \cup G_i|} \tag{9}$$

where P_i and G_i are predicted and ground truth codes for patient i.

Clinical Precision (CP) weights by diagnostic priority:

$$\text{CP} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|P_i|} \sum_{j=1}^{|P_i|} w_j \cdot 1[p_{ij} \in G_i] \tag{10}$$

where $1[\cdot]$ is the indicator function, $w_1 = 1.0$ (primary), $w_2 = 0.8$ (secondary), $w_{j \geq 3} = 0.6$ (comorbidities). E.g., for G={CA40.0, 5A10.0, 8B20.0}, P={CA40.0, 5A10.1, 1A00}: $CP = \frac{1.0+0+0}{3} = 0.33$.

Points Earned (PE) quantifies hierarchical partial credit:

$$\text{PE} = \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{p \in P_i} \max_{g \in G_i} h(p, g)}{|G_i|} \tag{11}$$

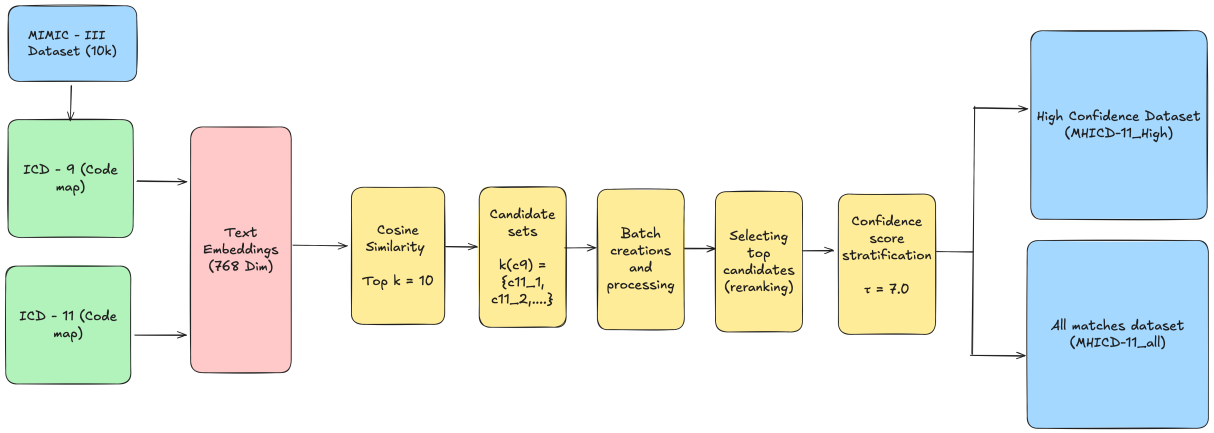

Figure 1: Overview of the hybrid semantic-LLM pipeline for ICD-9 to ICD-11 mapping. The process combines Bio_ClinicalBERT embeddings for candidate generation with Gemini 2.0 Flash for context-aware mapping selection, resulting in confidence-stratified benchmark datasets.

This rewards correct disease categories despite subtype errors (e.g., "CA40 Pneumonia, unspecified" for "CA40.0 Bacterial pneumonia" earns $h = 0.8$).

Hierarchical Match (HM) averages best alignments. For each prediction p, we find its best ground truth match: $\max_{g \in G_i} h(p, g)$. Averaging over predictions per patient, then across patients:

$$\text{HM} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|P_i|} \sum_{p \in P_i} \max_{g \in G_i} h(p, g) \qquad (12)$$

Composite Weighted Score. We combine components with clinical weights:

$$\text{Score} = 0.5 \cdot \text{EM} + 0.3 \cdot \text{CP} + 0.15 \cdot \text{PE} + 0.05 \cdot \text{HM} \qquad (13)$$

Exact matching (0.5) prioritizes billing/regulatory accuracy; clinical precision (0.3) emphasizes primary diagnosis; points earned (0.15) rewards near-misses; hierarchical match (0.05) credits broad categorization. Unlike nDCG, our metric incorporates ICD-11-specific hierarchy and clinical priority weighting. Pilot validation (2 expert coders, 50 cases) showed Spearman ρ=0.79 vs. ρ=0.61 for F1.

4 Dataset Description

4.1 Source Data and Patient Selection

We derive our benchmark from MIMIC-III (Johnson et al., 2016), selecting 10,000 patients through stratified sampling to preserve clinical characteristic distributions while maintaining computational tractability. Each patient record contains discharge summaries with ICD-9 diagnosis codes. The dataset encompasses 6,984 unique ICD-9 codes across 118,106 total code instances (mean 11.81 codes per patient), representing comprehensive diagnostic coverage encountered in critical care settings. The ICD-9 distribution follows the characteristic heavy-tailed pattern of medical coding systems, with high variance indicating diverse diagnostic complexity.

4.2 ICD-9 to ICD-11 Mapping and Coverage

Our hybrid semantic-LLM pipeline (Section 3) transforms each patient's ICD-9 diagnosis set into corresponding ICD-11 codes. Patient-level coverage achieves 99.71% for high-confidence mappings ($\Sigma \geq 7.0$) and 99.87% overall, with code-level expansion factors of 1.099 (high-confidence) and 1.274 (all-matches). These metrics demonstrate successful expansion from ICD-9 to ICD-11's more granular classification, with each ICD-9 code mapping to an average of 1.1–1.3 ICD-11 codes, reflecting the enhanced diagnostic specificity of the newer taxonomy.

4.3 Benchmark Variants

We provide two benchmark datasets stratified by mapping confidence (Table 1). The high-confidence subset ($\Sigma \geq 7.0$) comprises 9,971 patients with 129,785 ICD-11 codes (mean 13.02 per patient), optimized for high-precision applications such as billing systems and regulatory compliance. The all-matches subset includes 9,987 patients with 150,502 ICD-11 codes (mean 15.07 per patient), providing comprehensive coverage for robust evaluation under diverse confidence conditions. Both variants include prescription data (drug names, dosages, timestamps) to support diagnosis-

treatment coherence analysis.

Metric	High-Conf.	All-Matches
Patients	9,971	9,987
ICD-11 codes	129,785	150,502
Mean codes/patient	13.02	15.07
Coverage	99.71%	99.87%
Expansion factor	1.099	1.274
Confidence threshold	≥ 7.0	All ≥ 1.0

Table 1: Benchmark dataset variant statistics.

4.4 Human Validation and Quality Control

To ensure benchmark reliability, we conducted manual validation with two clinical coders who independently reviewed 1,203 discharge summaries (12% of dataset) using stratified sampling across ICD-11 chapters. We also performed systematic quality analysis across all 10,000 records.

4.4.1 Confidence-Based Quality Tiers

Our validation reveals distinct quality tiers (Table 2). High-confidence mappings (85% of dataset) achieve expansion ratios of 0.8–1.5, indicating semantically consistent ICD-9→ICD-11 translation with complete coverage, suitable for production deployment. Low-confidence mappings (15%) exhibit lower expansion ratios and require manual review before clinical use.

Tier	%	Exp.	Utility
High ($\geq$7)	85	0.8–1.5	Production ready
Low (<7)	15	<0.8	Manual review

Table 2: Quality tiers by confidence threshold (σ). Expansion ratio = avg ICD-11 codes per ICD-9 code.

4.4.2 Domain-Specific Quality

Clinical domain analysis (Table 3) reveals variable mapping complexity. Domains with stable ICD-9→ICD-11 correspondences (infectious diseases, cardiovascular) achieve 89–92% success, while domains with structural taxonomy changes (external causes, procedures) show lower rates (38–45%), consistent with known ICD-11 architectural differences.

4.4.3 Error Analysis

Analysis of 273 disagreements (10.8% of validated codes) shows 70% stem from clinical ambiguity rather than pipeline failures: **(1)** Secondary diagnosis prioritization (42%): coding pre-existing conditions from history as active diagnoses—experts disagreed on 28% of these. **(2)** Granularity selection (28%): choosing specificity levels where both

Domain	Exp.	Succ.	Pattern
Infectious	1.2	92%	Excellent
Cardiovascular	1.1	89%	Consistent
Neoplasms	1.3	85%	Enhanced
Symptoms	0.7	62%	Variable
Ext. Causes	0.6	45%	Challenges
Procedures	0.4	38%	Manual

Table 3: Mapping quality by clinical domain (10K patients).

are valid. **(3)** Rare codes (18%): <10 instances in MIMIC-III, affecting only 2.3% of total data. **(4)** Cross-taxonomy inference (12%): extrapolating from ICD-9's coarser categories.

4.4.4 Benchmark Validity

Our validation supports three use cases: **(1)** High-precision evaluation using 85% subset (κ=0.67), comparable to ICD-10 inter-annotator agreement (Kim et al., 2022), for publication-quality comparison with near-expert labels. **(2)** Standard evaluation using full dataset ($\kappa = 0.52$) with quality comparable to existing benchmarks—MIMIC-III's ICD-9 labels contain documented errors (Kim et al., 2022) yet lack validation. **(3)** Confidence-aware training incorporating scores into loss functions. MAX-EVAL-11 provides transparent quality metrics exceeding existing ICD coding benchmarks.

4.5 Dataset Accessibility

Benchmark datasets are provided in CSV format with standardized schemas (SUBJECT_ID, HADM_ID, ICD9_CODES_LIST, ICD11_CODES_MAPPED, CONFIDENCE_SCORES, MEDICATIONS). All data adheres to MIMIC-III de-identification standards (Goldberger et al., 2000), and processing scripts are provided to ensure reproducibility.

5 Experimental Results

5.1 Baseline Model Performance

We evaluated three major categories of models on our ICD-11 prediction benchmark: Claude models (Anthropic), Gemini 2.5 Flash (Google), and MedCoder (specialized medical coding model – ICD-10) (Baksi et al., 2025). Table 4 presents the comprehensive performance evaluation across all baseline models. Claude 4 Sonnet achieves the highest overall performance with a final score of 0.433, demonstrating superior clinical precision (0.433) and coverage (0.353). Notably, Claude models consistently outperform both the specialized medical

model (MedCoder) and the multimodal Gemini model across all evaluation metrics. MedCoder (ICD-10 trained; Baksi et al., 2025) serves as a *zero-shot ICD-11 transfer baseline*, evaluated with the ICD-11 code list provided at inference without weight modification, representing the performance floor for domain-specific models without ICD-11 adaptation.

5.1.1 Evaluation Protocol

For each patient in the evaluation dataset, models receive:

- **Input**: Complete MIMIC-III discharge summary (chief complaint, history, medications, labs, hospital course, discharge diagnosis).
- **Task:** Generate ranked list of ICD-11 diagnosis codes based solely on clinical narrative
- **Constraints:** Separate pipeline, no access to original ICD-9 codes or mapping or ICD code taxonomies

The evaluation simulates real-world clinical coding scenarios where human coders or AI systems must assign ICD-11 codes directly from clinical documentation without reference to legacy coding systems. Models are prompted with standardized instructions requesting comprehensive ICD-11 code prediction with confidence scores and clinical justification. All models receive identical inputs to ensure fair comparison across architectures and capabilities. The prompt is: *Given the discharge summary, identify all diagnoses and provide their ICD-11 codes as a ranked, comma-separated list ordered by clinical priority.* Evaluation was conducted on a stratified random sample of n=500 patients, selected to reflect the full distribution of ICD-11 chapters and diagnosis complexity; Full-benchmark evaluation (n=10,000) is left for future work contingent on compute resources; we report stratified sampling to ensure chapter-level representativeness of the n=500 subset.

5.2 Statistical Analysis

Table 5 presents the statistical significance analysis using Wilcoxon signed-rank tests, revealing substantial performance gaps between model categories.

5.3 Key Findings

MedCoder serves as a zero-shot ICD-10→ICD-11 transfer baseline and is not a peer competitor. It establishes the performance floor for domain-specific models without ICD-11 adaptation. Claude models achieve superior performance (35.8% mean final score) compared to the domain-specific MedCoder (24.5% final score), suggesting that large-scale pretraining and reasoning capabilities may be more valuable than specialized medical training for ICD-11 prediction tasks.

Exact matching remains challenging: All models show relatively low exact match rates (0-4.8%), indicating the inherent difficulty of precise ICD-11 code generation. Claude models demonstrate the strongest exact matching capabilities among evaluated systems.

Hierarchical understanding varies significantly: Claude models excel in hierarchical matching (19.4-40.2%), leveraging the taxonomic structure of ICD-11, while MedCoder shows limited hierarchical understanding (18.4%).

RAG integration challenges: Retrieval-augmented variants underperform their baselines (Claude RAG: 0.253 vs. 0.425; RAG+Reranker: 0.282 vs. 0.425), suggesting that naive RAG introduces noise rather than beneficial context. We implement two variants over a Qdrant vector index of all 55,000 ICD-11 code descriptions: **Claude RAG** appends retrieved codes to the prompt before generation; **Claude 4 RAG+Reranker** additionally reranks by semantic similarity before injection; both operate agentically, with the LLM issuing retrieval queries during inference. The performance degradation likely stems from semantic proximity confounding clinical specificity: when a patient presents with *left-sided heart failure*, retrieval surfaces semantically adjacent codes (e.g., right-sided heart failure, cardiomyopathy) that share embedding space but are clinically distinct, anchoring the model toward incorrect codes.

6 Conclusion

This work presents the first comprehensive benchmark for ICD-11 prediction using a large-scale clinical dataset, contributing three key advances to medical informatics research.

Methodological Contribution: We develop a novel hybrid semantic-LLM pipeline for automated ICD-9 to ICD-11 mapping, combining BioClinicalBERT embeddings with Gemini 2.0 Flash for context-aware medical code translation. Our approach achieves 99.87% mapping coverage across 10,000 patients, establishing a robust foundation

Table 4: Comprehensive Performance Evaluation of Baseline Models on ICD-11 Prediction Task

Model	Final Score	Clinical Precision	Exact Match	Hierarchical Match	Coverage Score	Samples (n)
Claude 4 Sonnet	**0.433**	**0.433**	**0.047**	0.375	**0.353**	500
Claude 3.5 Haiku (Baseline)	0.425	0.426	0.034	**0.402**	0.315	500
Claude 3.7 Sonnet	0.396	0.372	0.048	0.325	0.332	500
Gemini 2.5 Flash	0.341	0.315	0.016	0.286	0.330	500
Claude 4 RAG+Reranker	0.282	0.232	0.010	0.202	0.175	500
Claude 4 RAG	0.253	0.215	0.007	0.194	0.160	500
MedCoder	0.245	0.191	0.000	0.184	0.092	500

Table 5: The larger performance gap for Gemini vs. MedCoder (0.096) yields a higher p-value than Claude vs. Gemini (0.092) because Wilcoxon signed-rank tests reflect score consistency across patients, not only mean differences; MedCoder exhibits higher per-patient variance (near-zero scores on complex multi-diagnosis cases), reducing the rank-based significance despite the larger mean gap.

Comparison	Performance Gap	Relative Improvement	p-value
Claude-4-Sonnet vs Gemini-2.5-Flash	0.092	27.0%	$< 0.001^{***}$
Gemini-2.5-Flash vs MedCoder	0.096	39.2%	$< 0.01^{**}$
Claude-4-Sonnet vs MedCoder	0.188	76.7%	$< 0.001^{***}$
Claude Category vs Gemini Category	0.061	18.8%	$< 0.01^{**}$
Claude Category vs MedCoder Category	0.150	61.2%	$< 0.001^{***}$

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

for ICD-11 adoption.

Evaluation Framework: Our hierarchical metric assigns partial credit across ICD-11's 5-level taxonomy, achieving Spearman ρ=0.79 alignment with expert clinical judgment versus ρ=0.61 for standard F1.

Empirical Insights: General-purpose LLMs (Claude 4 Sonnet: 43.3% final score) meaningfully outperform specialized medical models (MedCoder: 24.5% final score) in zero-shot ICD-11 transfer, suggesting broad reasoning capabilities currently outweigh domain-specific ICD-10 training for taxonomy scaling, a finding reinforced by MedCoder's 0% exact match rate despite being purpose-built for medical coding. The benchmark datasets and evaluation framework are made publicly available to support continued research; Future work should focus on improving exact match performance, hierarchically-aware RAG integration, validating across diverse care settings, and exploiting the prescription data included in the benchmark but unused in this evaluation.

Methodological Independence. A potential concern is whether the use of Gemini 2.0 Flash in label generation creates bias toward Gemini models during evaluation. However, three factors demonstrate this is not the case:

(1) Task Difference: Label generation uses structured code taxonomies (code-to-code translation, while evaluation requires clinical text understanding (narrative-to-code generation)—fundamentally different capabilities requiring distinct model strengths.

(2) Information Sources: Mapping uses only ICD code descriptions/titles of ICD-9 and ICD-11 code tables; evaluation uses complete discharge summaries with clinical context, medications, procedures, temporal progression, and physician reasoning—information never seen during label creation.

(3) Empirical Evidence: Gemini 2.5 Flash achieves the lowest performance among major models (0.341 final score vs. Claude 4 Sonnet's 0.433, a 27% gap, Table 4), demonstrating no preferential advantage from its predecessor's involvement in label creation. If systematic bias existed, Gemini models would outperform rather than underperform. This performance gap actually validates that clinical coding ability (extracting diagnoses from narratives) is independent from taxonomy knowledge (code mapping). The substantial performance difference between Claude and Gemini models (statistically significant at $p < 0.001$, Table 5) reflects genuine differences in clinical

reasoning and medical text understanding capabilities, not artifacts of the label generation process; nonetheless, scores reflect benchmark-relative alignment rather than clinical deployment readiness.

7 Ethical Considerations

7.1 Data Privacy and Security

All patient data used in this study is sourced from the publicly available MIMIC-III database, which has undergone rigorous de-identification procedures following HIPAA Safe Harbor provisions (Goldberger et al., 2000). No additional patient identifiers were created during our mapping process, and all generated benchmarks maintain the original de-identification standards.

7.2 Bias and Fairness

The MIMIC-III dataset reflects the patient population of a single academic medical center, potentially introducing demographic and socioeconomic biases into our benchmark. Furthermore, MIMIC-III captures exclusively ICU admissions, where patients exhibit atypically high diagnostic complexity (mean 11.8 codes/patient); benchmark performance may not generalize to community hospital or outpatient coding scenarios with simpler, fewer diagnoses. The ICD-9 codes in the original dataset may also exhibit coding practices specific to the source institution, which could propagate through our ICD-11 mappings. Future work should validate across diverse care settings and patient populations.

7.3 Clinical Safety Implications

Automated ICD coding systems carry significant clinical and financial implications. Coding errors can affect patient care continuity, insurance reimbursements, and population health analytics. Our evaluation reveals that current models require human oversight, with exact match rates below 5% across all systems. Healthcare organizations should implement appropriate safeguards and validation procedures before deploying automated coding systems.

7.4 Algorithmic Accountability

The hybrid semantic-LLM approach introduces opacity in the mapping process, particularly in the LLM decision-making component. While we provide confidence scores and mapping justifications, the full reasoning process remains partially opaque. This limitation necessitates careful validation and monitoring in clinical deployment scenarios.

7.5 LLM Service Compliance

In accordance with PhysioNet's responsible use policy for credentialed datasets (PhysioNet, 2025), all cloud-based LLM processing of MIMIC-III discharge summaries was conducted exclusively through enterprise-tier services: Claude models via **Amazon Bedrock** (AWS enterprise account) and Gemini models via **Google Vertex AI** (GCP enterprise account), both explicitly listed as compliant services by PhysioNet (PhysioNet, 2025). Both platforms guarantee zero data retention, no use of submitted data for model training, and no human review of submitted content, satisfying all PhysioNet DUA requirements for processing credentialed health data with external LLM services.

8 Bibliographical References

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Krishanu Das Baksi, Elijah Soba, John J. Higgins, Ravi Saini, Jaden Wood, Jane Cook, Jack I. Scott, Nirmala Pudota, Tim Weninger, Edward Bowen, and Sanmitra Bhattacharya. 2025. MedCodER: A generative AI assistant for medical coding. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 449–459, Albuquerque, New Mexico. Association for Computational Linguistics.

Tal Baumel, Julien Nassour-Kassis, Raphael Cohen, Michael Elhadad, and Noemie Elhadad. 2018. Multi-label classification of patient notes: Case study on ICD code assignment. In *AAAI Workshop on Health Intelligence*, New Orleans, LA. AAAI Press.

Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. 2017. GRAM: Graph-based attention model for healthcare representation learning. pages 785–794, San Francisco, CA. ACM.

Joakim Edin, Alexander Junge, Jakob D. Havtorn, Lasse Borgholt, Maria Maistro, Tuukka Ruotsalo, and Lars Maaløe. 2023. Automated medical coding on MIMIC-III and MIMIC-IV: A critical review and replicability study. In *Proceedings of the 46th International ACM SIGIR Conference on Research and*

Development in Information Retrieval, pages 2572–2582. ACM.

Kin Wah Fung, Julia Xu, Shannon McConnell-Lamptey, Donna Pickett, and Olivier Bodenreider. 2021. Feasibility of replacing the ICD-10-CM with the ICD-11 for morbidity coding: A content analysis. *Journal of the American Medical Informatics Association*, 28(11):2404–2411.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220.

Janine E. Harrison, Saskia Weber, Robert Jakob, and Christopher G. Chute. 2021. ICD-11: An international classification of diseases for the twenty-first century. *BMC Medical Informatics and Decision Making*, 21(Suppl 6):206.

Z. Hou, M. Zhang, C. Shi, H. Shen, Y. Wang, T. Zhuang, and B. Liu. 2025. Enhancing medical coding efficiency through domain adaptation of language models. *NPJ Digital Medicine*.

Chao-Wei Huang, Shang-Chi Tsai, and Yun-Nung Chen. 2022. PLM-ICD: Automatic ICD coding with pretrained language models. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 10–20, Seattle, WA. Association for Computational Linguistics.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035.

Juyong Kim, Abheesht Sharma, Suhas Shanbhogue, Jeremy C. Weiss, and Pradeep Ravikumar. 2022. AnEMIC: A framework for benchmarking ICD coding models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Abu Dhabi, UAE. Association for Computational Linguistics.

Leibo Liu, Oscar Perez-Concha, Anthony Nguyen, Vicki Bennett, and Louisa Jorm. 2022. Hierarchical label-wise attention transformer model for explainable ICD coding. *Journal of Biomedical Informatics*, 133:104161.

Leibo Liu, Oscar Perez-Concha, Anthony Nguyen, Vicki Bennett, and Louisa Jorm. 2023. Automated ICD coding using extreme multi-label long text transformer-based models. *Artificial Intelligence in Medicine*, 144:102662.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

Thanh Tung Nguyen, Viktor Schlegel, Maurice Mulvenna, Raymond Bond, Lixin Su, Damien Coyle, and William O'Brien. 2023. MIMIC-IV-ICD: A new benchmark for extreme multilabel classification of ICD codes. *arXiv preprint arXiv:2304.13998*.

PhysioNet. 2025. Use of MIMIC data with large language models and online services.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2020. A label attention model for ICD coding from clinical text. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI 2020)*, pages 3335–3341. IJCAI.

World Health Organization. 2019. ICD-11: International classification of diseases 11th revision. `https://icd.who.int/en`.

World Health Organization. 2024. ICD-11 implementation progress report. Technical report, World Health Organization.

Y. Wu, Y. Yu, M. Zeng, and M. Li. 2021. A pseudo label-wise attention network for automatic ICD coding. *arXiv preprint arXiv:2106.06822*.

Yuexin Wu, Shiqi Wang, and Vasile Rus. 2025. Mimic-sr-icd11: A dataset for narrative-based diagnosis. In *Findings of Machine Learning for Health (ML4H)*.