Multi-Model Hierarchical Semantic Retrieval for Predicting Abbreviated Injury Scale 2015 from Clinical Narratives

Chien-Ming Lee*, Pei-Ling Lee†, Chao-Wen Chen†, Joffrey Hsu†, and Chia-Yeuan Han*

*Department of Medical Information,
Kaohsiung Medical University Hospital, Kaohsiung Medical University,
Kaohsiung, Taiwan
nomoney.lee@gmail.com, hoganhan2@gmail.com

†Division of Trauma and Surgical Critical Care, Department of Surgery,
Kaohsiung Medical University Hospital, Kaohsiung Medical University,
Kaohsiung, Taiwan
peiling@trauma.tw, kmutrauma@gmail.com, joffrey.hsu@gmail.com

Abstract - **Accurate Abbreviated Injury Scale coding is pivotal for trauma registries but remains time-consuming due to the complexity of unstructured clinical narratives. To automate this process while balancing efficiency and accuracy, a hierarchical semantic routing framework is proposed that integrates Bi-Encoder and Cross-Encoder architectures. Unlike traditional brute-force approaches, the proposed method employs a confidence-based routing strategy: (1) a lightweight Bi-Encoder first retrieves candidate codes based on vector similarity; (2) if the prediction confidence exceeds a calibrated threshold (τ_h), the result is accepted immediately to conserve resources; (3) for ambiguous or low-confidence cases, a fallback mechanism triggers the computationally intensive Cross-Encoder to re-rank the top candidates. This design effectively acts as a semantic noise filter, leveraging the Bi-Encoder to prune the search space and the Cross-Encoder to resolve fine-grained distinctions. The system was evaluated using a real-world dataset of emergency department narratives. Experimental results demonstrate that the proposed method achieves a Top-1 accuracy of 59.3%, notably outperforming the full Cross-Encoder baseline (56.5%), while delivering over 14× speedup in inference time (547 ms/sample). The main output of this paper is a practical, human-in-the-loop system that operates efficiently on standard central processing unit hardware to expedite injury coding and reduce the cognitive burden on trauma registrars.**

Keywords - abbreviated injury scale; natural language processing; hierarchical classification; semantic retrieval.

I. Introduction

This paper presents a comprehensive extension of the preliminary research [1], aiming to address the persistent challenges in automating injury severity coding. Traumatic injury remains a leading cause of mortality and morbidity worldwide, necessitating rigorous data collection to drive quality improvement and epidemiological research. In this context, the Abbreviated Injury Scale (AIS) 2015 [2] serves as the globally recognized standard for classifying injury severity. Accurate AIS coding is the fundamental prerequisite for calculating the Injury Severity Score (ISS), which is critical for predicting patient survival, benchmarking hospital performance, and allocating healthcare resources.

However, the AIS coding process is inherently complex. The AIS 2015 dictionary contains over 2,000 distinct codes, each associated with a specific anatomical description, severity level, and body region. Given this intricate hierarchical structure and extensive coverage, manual coding demands high clinical expertise and rigorous training. Trauma registrars must mentally map unstructured clinical observations to precise dictionary entries—a process that is labor-intensive, time-consuming, and prone to inter-rater variability. Consequently, developing an automated auxiliary system to assist registrars has become an imperative task in medical informatics, promising to enhance both coding efficiency and data consistency.

The primary obstacle to automation lies in the inherent nature of the input data. Diagnoses documented by clinicians in emergency rooms are typically unstructured, free-text narratives replete with abbreviations (e.g., ”fx” for fracture), colloquialisms, non-standard spellings, and semantic ambiguity. For instance, a ”crush injury to the chest” implies a completely different set of AIS codes compared to a ”penetrating chest wound,” requiring a system with deep semantic understanding to distinguish these nuances effectively.

Challenges in Large-Scale Semantic Retrieval Beyond the linguistic noise, a significant technical challenge lies in the **trade-off between retrieval accuracy and computational efficiency.** Traditional keyword-based methods fail to capture semantic nuances, while modern deep learning approaches face a scalability bottleneck. A standard ”Bi-Encoder” approach is fast but often lacks the granularity to distinguish between highly similar anatomical descriptions (e.g., specific vertebrae levels). Conversely, a ”Cross-Encoder” approach, which performs full self-attention over input pairs, offers superior accuracy but is computationally prohibitive when re-ranking thousands of dictionary candidates for every patient record. Solving this ”accuracy-efficiency dilemma” without

relying on expensive GPU infrastructure remains an open problem in clinical Natural Language Processing (NLP) deployment.

Research Gap and Motivation As identified in the preliminary work [1], previous attempts utilizing basic keyword mapping and standard pre-trained models showed potential but faced significant limitations. That approach adopted a relatively simplistic architecture that did not fully exploit the hierarchical taxonomy of the AIS dictionary, nor did it effectively address the inference latency issues inherent in large-scale retrieval. Accuracy and reliability in handling complex semantic ambiguity remained insufficient for clinical standards. Consequently, there is a compelling need to develop a new architecture that is more accurate, robust, and capable of mimicking the hierarchical reasoning process of human experts while maintaining operational efficiency.

Proposed Framework and Contributions To overcome these challenges and significantly improve the performance and practicality of AIS 2015 coding, this paper proposes a **Multi-Model Hierarchical Semantic Retrieval Framework**. Unlike monolithic models, the proposed approach mimics the cognitive process of a human coder: first narrowing down the anatomical region and then scrutinizing the specific injury details. The framework integrates a lightweight Bi-Encoder for rapid candidate generation and a high-precision Cross-Encoder for reranking difficult cases. Crucially, the framework proposed in this paper operates as a training-free retrieval system. By leveraging publicly pre-trained biomedical language models without the need for task-specific fine-tuning, the dependency on large-scale annotated datasets is eliminated, thereby addressing the data scarcity bottleneck and enabling immediate deployment in hospitals lacking historical labeled records.

The main contributions of this paper are summarized as follows:

- **A Novel Hierarchical Routing Architecture**: A multi-stage retrieval framework is proposed that dynamically routes input narratives based on prediction confidence. This design utilizes a similarity-based classifier (Bi-Encoder) to accurately prune the search space and a Cross-Encoder to handle fine-grained semantic comparisons, effectively balancing system performance.
- **Efficiency-Accuracy Optimization**: The proposed framework is demonstrated to act as a semantic noise filter. By reserving the computationally expensive Cross-Encoder only for ambiguous inputs (the "long-tail" cases), the proposed method achieves a $\sim 14\times$ speedup compared to a full Cross-Encoder baseline while paradoxically achieving higher Top-1 accuracy by filtering out irrelevant distractors.
- **Practical Deployment Feasibility**: Unlike many academic models that require high-end GPUs, the proposed optimized pipeline is designed to run efficiently on standard CPU hardware (inference time $\approx$ 547 ms), making it a viable decision-support tool for resource-constrained hospital environments.
- **Comprehensive Evaluation**: The proposed method is validated on a real-world dataset of emergency department narratives, providing a detailed analysis of performance across different anatomical regions and discussing the implications for human-in-the-loop clinical workflows.

The remainder of the paper is organized as follows: Section II reviews the related work in medical coding and semantic retrieval. Section III details the hierarchical semantic retrieval procedures and model architectures. Section IV presents the experimental setup and evaluation results. Section V discusses the findings, the mechanism of efficiency gains, and limitations. Finally, Section VI concludes the paper and outlines future directions.

II. Related Work

This section reviews related research in the domain of automated medical coding, with a particular focus on processing unstructured clinical narratives, leveraging ontological structures, and adoption of multi-stage retrieval frameworks. Collectively, these advancements lay the foundation for the Multi-Model Hierarchical Semantic Retrieval Framework proposed in this paper.

The review is organized as follows: Section II-A highlights the fundamental challenges in automated medical coding and the rise of deep learning applications. Section II-B discusses strategies for handling the structured, hierarchical nature of medical coding systems. Section II-C indicates challenges of LLMs in Medical Coding. Finally, Section II-D surveys recent advancements in multi-model retrieval and re-ranking mechanisms relevant to proposed work.

A. Challenges in Automated Medical Coding and Deep Learning Applications

According to recent studies [3] and [4], Automated Medical Coding (AMC), which involves assigning International Classification of Diseases (ICD) codes to clinical documents, is a pivotal task in healthcare operations and service delivery. Given that manual coding is time-consuming and error-prone, researchers have widely applied deep learning and NLP techniques to automate this process.

Ji et al. [3] identified that the AMC task faces several primary challenges:

1) **Lengthy and Noisy Text**: Clinical notes typically consist of long, free-text replete with medical jargon, non-standard abbreviations, and semantic ambiguity.
2) **High-dimensional Label Space**: Medical ontologies (e.g., ICD-9/10) contain tens of thousands of codes, rendering the task an extreme multi-label classification problem.
3) **Class Imbalance**: The distribution of codes is highly skewed, exhibiting a "long-tail phenomenon" involving common and rare diseases.

Mullenbach et al. [4] noted that early deep learning approaches extensively utilized Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs). Notably,

CNN architectures, such as CAML (Convolutional Attention for Multi-Label classification), combine convolutional layers with attention mechanisms to extract specific segments from the document relevant to each code, thereby providing interpretable predictions. While these studies primarily focus on ICD coding, the tasks of AIS coding share these identical challenges, with the added complexity of mapping injuries to specific severities.

However, a critical limitation of these aforementioned deep learning architectures is their heavy reliance on the **supervised learning** paradigm. To achieve optimal performance, models like CAML require large-scale, high-quality annotated corpora to learn the complex mapping between clinical narratives and code definitions. This dependency poses a significant bottleneck for trauma registries, where labeled historical data is often scarce, expensive to obtain, or subject to privacy restrictions.

To address this limitation, the proposed work diverges from the traditional supervised approach and adopts a **training-free retrieval** strategy. By leveraging the inherent semantic knowledge embedded in publicly pre-trained language models, the proposed framework circumvents the need for task-specific training data. This paradigm shift not only eliminates the costs associated with data annotation but also significantly lowers the barrier to deployment.

B. Handling the Structured Nature of Medical Coding

Analogous to ICD codes, the AIS 2015 lexicon utilized in this study possesses an inherent hierarchical structure [5][6]. Previous work [6] has demonstrated that explicitly modeling the code hierarchy and correlations is pivotal for enhancing coding accuracy.

1) **Hierarchical Classification**: Recent deep learning models have integrated hierarchical information by designing specialized architectures. For instance, Cao et al. [6] proposed a hyperbolic representation method that utilizes the tree-like properties of hyperbolic space to learn hyperbolic representations of the code hierarchy, while incorporating Graph Convolutional Networks (GCNs) to capture code co-occurrence relationships. Furthermore, the hierarchical joint learning mechanism proposed by Vu et al. [7] leverages predictions of the first three characters of ICD codes to assist in predicting complete codes, thereby improving performance on low-frequency codes. Inspired by these hierarchical approaches, the proposed strategy (Phase 1: AIS root node classification) focuses on identifying the semantically closest root node, effectively pruning the search space to enhance overall coding accuracy.
2) **Semantic Representation and Entity Alignment**: Given the complexity of clinical text, establishing robust medical entity representations is critical. Domain-specific pre-trained models, such as BioBERT [8], have been shown to enhance the understanding of in-domain text through pre-training on biomedical corpora. Concurrently, to address the high heterogeneity of medical entity naming, research has focused on learning entity representations. For example, the BIOSYN model [9] achieves state-of-the-art performance in biomedical entity normalization tasks by employing synonym marginalization techniques. In this work, a SapBERT model [10] is employed in the first phase. Distinct from standard models, SapBERT is specifically designed with a self-alignment pre-training scheme. By leveraging synonym relationships within large-scale medical ontologies such as the Unified Medical Language System (UMLS), it pulls entity names representing the same concept closer together, forming compact clusters in the vector space. This enables SapBERT to efficiently capture the nuances of medical terminology, making it highly suitable for the semantic similarity classification task in this study.

C. Generative LLMs versus Discriminative Retrieval in Medical Coding

The advent of Large Language Models (LLMs), such as GPT-4 and Llama-3, has transformed biomedical NLP, demonstrating remarkable zero-shot generalization capabilities [11]. Recent studies have explored their use as classifiers for automated ICD coding [12], suggesting that generative models could, in principle, directly predict AIS codes from clinical narratives.

However, deploying LLMs in high-stakes medical registry tasks poses substantial challenges:

1) **Hallucination Risks**: Generative models can produce 'medical hallucinations' [13], particularly for rare pathologies or under instruction-alignment bias. In contrast, the proposed discriminative retrieval framework constrains outputs strictly to the AIS dictionary, inherently preventing the generation of non-existent codes.
2) **Resource and Privacy Constraints**: LLM inference demands significant GPU memory and computational resources, often requiring cloud-based services that raise data privacy concerns. By design, the proposed lightweight BERT-based framework runs efficiently on standard CPU hardware, enabling secure, on-premise deployment in hospital environments.
3) **Confidence Calibration**: While LLMs generate fluent text, reliably quantifying prediction confidence is challenging. The proposed hierarchical routing mechanism leverages explicit similarity scores to determine when fallback re-ranking is necessary, providing transparent and calibrated uncertainty measures essential for human-in-the-loop workflows.

Therefore, although LLMs represent a promising direction in biomedical NLP, for trauma coding applications that demand verifiability, robustness, and efficiency, a specialized, retrieval-based architecture remains the preferred solution.

D. Multi-model Retrieval and Re-ranking Mechanisms

The second and third phases of the proposed framework involve fine-grained semantic comparison and a fallback mechanism under low-confidence conditions. This design aligns with the Retrieve-and-Re-rank paradigm widely adopted in the field of Information Retrieval (IR) [14].

1) **Efficient Semantic Retrieval (Bi-Encoders)**: To achieve efficient semantic comparison, models must overcome the computational limitations associated with Cross-Encoder architectures like BERT. Sentence-BERT (SBERT) [15] introduces a Bi-Encoder architecture that modifies BERT by using Siamese and triplet network structures to derive semantically meaningful sentence embeddings. SBERT reduces the computational overhead of similarity search from tens of hours to mere seconds by enabling cosine similarity comparison, making it highly suitable for large-scale semantic retrieval. In the global semantic retrieval phase, it is leveraged that the Sentence-Transformers framework [15] initialized with BioBERT weights [16] to generate high-dimensional semantic embedding vectors.
2) **Contextual Re-ranking and Cross-Attention (Cross-Encoders)**: While bi-encoders are efficient, their performance is typically inferior to cross-encoders, as the latter can perform full Cross-Attention on input pairs [17]. BERT cross-encoders calculate relevance scores by processing the query and document simultaneously [18], [19], thereby achieving state-of-the-art accuracy in re-ranking tasks. This study deploys a Cross-Encoder model [20] within the fallback mechanism. This allows for precise, fine-grained scoring on the candidate pool generated by the retrieval phase, ensuring high accuracy in the final output under low-confidence conditions.

In summary, the Multi-Model Hierarchical Semantic Retrieval Framework proposed in this study integrates specialized techniques designed to address the unique challenges of medical coding—specifically hierarchy and semantic complexity—within an established Retrieve-and-Rerank architecture from the field of Information Retrieval (IR). By synergizing the precise domain entity alignment capabilities of SapBERT, the efficient semantic retrieval speed of SBERT, and the high-precision re-ranking performance of Cross-Encoders, this approach delivers a solution for AIS 2015 coding that is simultaneously accurate, robust, and efficient.

III. Proposed Method

This section details the architecture and computational workflow of the proposed system.

A. Overview

In this study, a Multi-Model Hierarchical Semantic Retrieval Framework is proposed to automate the prediction of AIS 2015 codes. Distinct from traditional supervised learning approaches that rely on extensive annotated datasets, the proposed framework operates as a *training-free semantic retrieval architecture*. By strictly leveraging the inherent generalization capabilities of pre-trained models (specifically SapBERT and BioBERT) within their original semantic spaces, this design eliminates the need for task-specific fine-tuning, thereby ensuring high deployability across diverse clinical settings without the bottleneck of labor-intensive data labeling.

The proposed framework operates in a hierarchical manner, leveraging the tree-like structure of the AIS dictionary. It consists of three distinct phases: (1) a Hierarchical Strategy for candidate narrowing, (2) Semantic Comparisons, and (3) a Robustness Enhancement Mechanism featuring a fallback global re-ranking module.

The selection of specific pre-trained models is strategically aligned with the unique requirements of each stage. In Stage 1, SapBERT is utilized for its specialized capability in biomedical entity alignment; its training objective focuses on clustering synonyms, making it highly effective for mapping diverse clinical narratives to the correct anatomical root nodes. In contrast, for Stage 2 and the fallback stage, BioBERT is employed to perform fine-grained matching. While SapBERT excels at individual entity representation, BioBERT offers a broader understanding of medical context and complex clinical syntax, which is crucial for distinguishing between highly similar injury descriptions within a specific anatomical subtree.

B. AIS Hierarchy and Root Extraction

The AIS 2015 is inherently hierarchical, organized by body regions, anatomical structures, and injury severity. To facilitate hierarchical retrieval, the AIS dictionary is first restructured into a forest of semantic trees. Let $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$ represent the set of all unique injury descriptions in the AIS dictionary. The top-level nodes (roots) corresponding to major body regions or specific anatomical chapters are identified. For each root r_k, the textual description $T(r_k)$ is extracted. This process partitions the search space into subsets S_k, where each subset contains injury descriptors belonging to the k-th semantic tree.

C. Text Embedding Method

Given an input clinical narrative q (e.g., a diagnosis string), the first objective is to identify the most relevant anatomical branch. A publicly pre-trained SapBERT model [21] is employed as the semantic encoder for this phase. SapBERT is optimized for biomedical entity alignment, making it highly effective at capturing the nuances of medical terminology. This process is implemented by employing a mean-pooling mechanism to aggregate the hidden states output by the encoder.

Formally, the input q is tokenized into a sequence of tokens $T = \{t_1, t_2, \ldots, t_L\}$. The encoder processes T to output a corresponding sequence of hidden states $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_L\}$. Since sequence lengths vary, the method uses an Attention Mask $\mathbf{M}$ for weighted averaging, ensuring that only the actual tokens are averaged to obtain a valid sentence embedding $\mathbf{e}$.

The specific calculation steps are as follows:

1) **Masked Summation**: The hidden states $\mathbf{H}$ are weighted by the attention mask $\mathbf{M} = \{m_1, \ldots, m_L\}$ and then

summed across the sequence length dimension to yield the masked sum $\mathbf{S}$:

$$\mathbf{S} = \sum_{i=1}^{L} m_i \mathbf{h}_i \tag{1}$$

where $m_i \in \{0, 1\}$ is the mask value corresponding to $\mathbf{h}_i$.

2) **Effective Token Count**: The total number of effective tokens C in the mask is calculated, applying a minimum value $\epsilon = 10^{-9}$ to prevent division by zero:

$$C = \max\left(\sum_{i=1}^{L} m_i, \epsilon\right) \tag{2}$$

3) **Mean Pooling**: The average embedding vector $\mathbf{e}_{mean}$ is computed as:

$$\mathbf{e}_{mean} = \frac{\mathbf{S}}{C} \tag{3}$$

4) $\boldsymbol{L_2}$ **Normalization**: Finally, the average embedding $\mathbf{e}_{mean}$ is subjected to L_2 normalization. This ensures all embedding vectors reside on the unit sphere, which is crucial for cosine similarity-based classification (as described below):

$$\mathbf{e} = \frac{\mathbf{e}_{mean}}{\|\mathbf{e}_{mean}\|_2} \tag{4}$$

D. Cosine Similarity-based Classification (Stage 1: AIS Root Classification)

This stage aims to assign an input diagnostic embedding $\mathbf{e}$ to the most relevant top-level AIS root nodes, thereby narrowing the search space for subsequent fine-grained classification. Rather than performing direct label prediction, this stage serves as a semantic routing mechanism that partitions the AIS dictionary into anatomically coherent subsets, as defined in the Method Overview.

Specifically, each AIS root r_k is associated with a textual description $T(r_k)$, from which a normalized, fixed-dimensional root embedding $\mathbf{r}_k$ is pre-computed using the same text embedding method described in the previous section. Given an input diagnostic embedding $\mathbf{e}$, the relevance between $\mathbf{e}$ and each root node is quantified using cosine similarity, enabling the identification of the most semantically aligned anatomical roots.

1) **Similarity Calculation**: Since both the input embedding $\mathbf{e}$ and each root embedding $\mathbf{r}_j$ are L_2-normalized, their inner product is equivalent to cosine similarity:

$$\text{Sim}_j = \mathbf{e}^\top \mathbf{r}_j = \cos(\mathbf{e}, \mathbf{r}_j) \tag{5}$$

Let $\mathbf{R} \in \mathbb{R}^{K \times d}$ be the matrix containing all root embeddings. The similarity scores are computed simultaneously in a vectorized manner as

$$\mathbf{Sim} = \mathbf{R}\mathbf{e}.$$

2) **Result Ranking and Output**: The similarity scores are ranked in descending order to produce a list of AIS root candidates along with their corresponding similarity scores. This ranked list is used to determine the candidate subsets for the subsequent classification stage.

E. Final AIS Code Refinement and Selection (Stage 2: AIS Code Refinement and Selection)

Given the candidate AIS root nodes identified in Stage 1, this stage focuses on selecting the most precise AIS injury code within the corresponding semantic subtree. By restricting the search space to a root-specific subset $\mathcal{S}_k$, the refinement process employs a higher-resolution semantic similarity model to perform fine-grained matching between the input diagnostic embedding and candidate AIS codes. Here, $\mathcal{S}_k \subset \mathcal{D}$ denotes the set of AIS injury descriptions associated with the selected root node r_k.

1) Similarity Thresholding and Subtree Pruning: Upon obtaining the similarity vector $\mathbf{Sim}$ from Stage 1, the system identifies the top-ranked root node r^* with the highest similarity score Sim_{max}. To ensure the reliability of the semantic routing, a predefined high-confidence threshold τ_h is applied. The pruning mechanism validates whether:

$$\text{Sim}_{max} \geq \tau_h \tag{6}$$

If this condition is met, r^* is accepted as the target anatomical branch. The system then effectively prunes the search space from thousands of potential codes to the specific semantic subset $\mathcal{S}_{r^*} \subset \mathcal{D}$ associated with r^*. This pruning step removes irrelevant anatomical regions, thereby reducing the search space and improving both computational efficiency and classification accuracy.

2) Semantic Similarity Calculation based on BioBERT: Within the constrained candidate subset $\mathcal{S}_{r^*}$, a pre-trained BioBERT model [22] is employed for fine-grained matching. To generate high-quality sentence embeddings, the model is implemented using the Sentence-Transformers [15] framework, transforming the clinical diagnosis and each candidate AIS description within $\mathcal{S}_{r^*}$ into high-dimensional semantic embeddings.

The calculation proceeds as follows:

1) **Vector Embedding**: Let q be the input clinical diagnosis and $\mathcal{C} = \{c_1, c_2, \ldots, c_M\}$ be the set of candidate AIS descriptions within the subset $\mathcal{S}_{r^*}$. The encoder function $E(\cdot)$ transforms these textual inputs into dense vectors:

$$\mathbf{v}_q = E(q), \quad \mathbf{V}_\mathcal{C} = \{E(c_1), E(c_2), \ldots, E(c_M)\} \tag{7}$$

where $\mathbf{v}_q$ is the diagnosis vector and $\mathbf{V}_{\mathcal{C}}$ is the matrix of candidate vectors.

2) **Cosine Similarity Calculation**: The matching score for each candidate is determined by calculating the cosine similarity between $\mathbf{v}_q$ and each candidate vector $\mathbf{v}_{c_i} \in \mathbf{V}_{\mathcal{C}}$:

$$\text{Score}_i = \cos(\mathbf{v}_q, \mathbf{v}_{c_i}) = \frac{\mathbf{v}_q \cdot \mathbf{v}_{c_i}}{\|\mathbf{v}_q\|_2 \|\mathbf{v}_{c_i}\|_2} \tag{8}$$

3) **Result Ranking and Output**: The system sorts all candidate AIS codes based on Score_i in descending order. The top K (e.g., $K = 10$) AIS codes with the highest scores are selected as the final output, providing a ranked list for user review.

F. Fallback Retrieval and Re-ranking Mechanism

This fallback mechanism is activated when the Stage 1 root node classifier fails to identify any root node whose similarity score exceeds the predefined threshold τ_h. Such cases typically arise when the input clinical diagnosis is highly ambiguous, contains sparse contextual cues, or is semantically distant from the existing AIS code descriptions.

To address these challenging scenarios, the system adopts a two-model retrieval architecture that performs a global semantic search across the entire AIS database, followed by a fine-grained re-ranking process. This design ensures robust and reliable AIS code retrieval even when hierarchical semantic routing is not sufficiently confident.

The first step of the fallback mechanism is the construction of a global Top-N candidate pool from the entire AIS code description set. This step performs a coarse-grained semantic retrieval to identify a manageable subset of potentially relevant AIS codes.

1) **Embedding Model**: A pre-trained BioBERT model [22] is used to encode the input clinical diagnosis q into a dense vector $\mathbf{v}_q$. The embedding vectors of all AIS code descriptions, denoted as $\mathbf{V}_{\text{AIS}}$, are pre-computed and stored offline to enable efficient retrieval.
2) **Global Similarity Calculation**: Cosine similarity scores are computed between $\mathbf{v}_q$ and all AIS code embeddings:

$$\mathbf{S}_{\text{all}} = \cos(\mathbf{v}_q, \mathbf{V}_{\text{AIS}}) \tag{9}$$

3) **Top-N Candidate Selection**: The system selects the top $N = 100$ AIS code descriptions with the highest similarity scores to form the candidate pool $\mathcal{P}_{100}$.

1) Cross-Encoder Re-ranking: To further improve matching precision, a pre-trained Cross-Encoder model [20] is applied to re-rank the candidates in $\mathcal{P}_{100}$ using fine-grained semantic interaction modeling.

1) **Paired Input Representation**: Unlike Bi-Encoder retrieval, the Cross-Encoder processes paired inputs of the form $[q, c_i]$, where $c_i \in \mathcal{P}_{100}$ denotes an AIS code description. This architecture jointly encodes both texts in a single forward pass, allowing the model to capture token-level interactions.
2) **Re-ranking Score Computation**: For each candidate AIS description $c_i \in \mathcal{P}_{100}$, the Cross-Encoder computes a relevance score by jointly encoding the clinical diagnosis q and the candidate text:

$$\text{Score}_i = f_{\text{CE}}(q, c_i), \quad \forall c_i \in \mathcal{P}_{100} \tag{10}$$

where $f_{\text{CE}}(\cdot, \cdot)$ denotes the scoring function implemented by the pre-trained Cross-Encoder model.
3) **Final Output Selection**: The candidate AIS codes are sorted in descending order based on the computed scores Score_i, and the top $K = 10$ codes are returned as the fallback prediction results.

This fallback strategy ensures robust AIS code retrieval by combining global semantic search with precise Cross-Encoder re-ranking, particularly in cases where hierarchical semantic routing is insufficiently confident.

IV. Results

This section presents the experimental results and evaluates the performance of the proposed framework.

A. Comparative Analysis with Baseline across Body Regions

Overall classification performance is first evaluated against the baseline method (BioBERT global retrieval). Table I details the Top-1, Top-5, and Top-10 accuracies across different AIS body regions.

The proposed method outperforms the baseline across the majority of anatomical regions, particularly achieving distinct advantages in soft-tissue and extremity categories. In complex areas such as the *Head*, *Neck*, and *Abdomen*, substantial accuracy gains are observed. For example, in the *Abdomen* region, Top-1 accuracy improved from **34.6%** (Baseline) to **80.8%** (Ours), highlighting the benefit of constraining the search space through semantic routing.

Performance in Spinal Regions: Performance across Spinal regions exhibited greater variability compared to the consistent gains seen in soft-tissue regions. In the **Cervical Spine**, the proposed method recorded a Top-1 accuracy of 25.0%, lower than the baseline's 31.3%. In the **Thoracic Spine**, while the Top-1 accuracy (29.4%) lagged, the Top-5 accuracy increased substantially to 64.7%, creating a gap of 35.3 percentage points between Top-1 and Top-5. The **Lumbar Spine** region showed improvements across all metrics, with Top-1 rising to 31.3% and Top-10 reaching 68.8%. These observations suggest that while the correct spinal codes are frequently retrieved within the top candidates, the final Top-1 ranking can be unstable across vertebral segments.

TABLE I. Performance comparison between the baseline (BioBERT global retrieval) and the proposed method across AIS body regions. Parentheses show the number of correct predictions and the total samples for each region.

Region (N)	**Method**	**Top-1 (%)**	**Top-5 (%)**	**Top-10 (%)**
Head (27)	Baseline	37.0 (10)	66.7 (18)	66.7 (18)
	Proposed	**44.4 (12)**	66.7 (18)	**77.8 (21)**
Face (26)	Baseline	50.0 (13)	69.2 (18)	**80.8 (21)**
	Proposed	**65.4 (17)**	**73.1 (19)**	73.1 (19)
Neck (26)	Baseline	46.2 (12)	69.2 (18)	88.5 (23)
	Proposed	**73.1 (19)**	**92.3 (24)**	**96.2 (25)**
Thorax (26)	Baseline	30.8 (8)	69.2 (18)	84.6 (22)
	Proposed	**57.7 (15)**	**92.3 (24)**	**92.3 (24)**
Abdomen (26)	Baseline	34.6 (9)	76.9 (20)	80.8 (21)
	Proposed	**80.8 (21)**	**96.2 (25)**	**96.2 (25)**
Cervical Spine (16)	Baseline	**31.3 (5)**	**75.0 (12)**	**81.3 (13)**
	Proposed	25.0 (4)	56.3 (9)	62.5 (10)
Thoracic Spine (17)	Baseline	**35.3 (6)**	52.9 (9)	52.9 (9)
	Proposed	29.4 (5)	**64.7 (11)**	**70.6 (12)**
Lumbar Spine (16)	Baseline	25.0 (4)	56.3 (9)	62.5 (10)
	Proposed	**31.3 (5)**	56.3 (9)	**68.8 (11)**
Other Spine (2)	Baseline	100.0 (2)	100.0 (2)	100.0 (2)
	Proposed	100.0 (2)	100.0 (2)	100.0 (2)
Upper Extremity (26)	Baseline	46.2 (12)	65.4 (17)	88.5 (23)
	Proposed	**69.2 (18)**	**88.5 (23)**	**92.3 (24)**
Lower Extremity (25)	Baseline	32.0 (8)	84.0 (21)	88.0 (22)
	Proposed	**68.0 (17)**	**92.0 (23)**	**92.0 (23)**
External (25)	Baseline	28.0 (7)	72.0 (18)	80.0 (20)
	Proposed	**44.0 (11)**	72.0 (18)	**84.0 (21)**
Other (27)	Baseline	59.3 (16)	88.9 (24)	**100.0 (27)**
	Proposed	**85.2 (23)**	**96.3 (26)**	96.3 (26)
All (285)	Baseline	39.3 (112)	71.6 (204)	81.1 (231)
	Proposed	**59.3 (169)**	**81.1 (231)**	**85.3 (243)**

Overall, despite minor performance trade-offs in certain spinal subregions, the proposed method achieves a superior overall Top-1 accuracy of **59.3%** compared to the baseline's **39.3%**, demonstrating the effectiveness of hierarchical semantic routing in improving both retrieval precision and coverage across diverse anatomical regions.

B. Effectiveness of the Hierarchical Routing Strategy

Next, the effectiveness of the proposed hierarchical routing strategy is analyzed. The core of the proposed system is the ability to distinguish between high-confidence samples (routed via the anatomical tree) and ambiguous samples (handled by the fallback mechanism). Table II presents the distribution of test samples and their respective classification accuracies.

TABLE II. Analysis of sample distribution and classification performance between the Hierarchical Routing (High Confidence) and Fallback Mechanism.

Mechanism	**Count** (N)	**Proportion**	**Top-1 (%)**	**Top-5 (%)**	**Top-10 (%)**
High Confidence (Stage 1 & 2)	39	13.7%	79.5 (31)	97.4 (38)	97.4 (38)
Fallback	246	86.3%	56.1 (138)	78.5 (193)	83.3 (205)
Total	285	100.0%	59.3 (169)	81.1 (231)	85.3 (243)

1) Performance of the High-Confidence Path: Approximately **13.7%** of the samples (39/285) were identified with high confidence (Sim $\geq \tau_h$). This subset achieved a remarkable Top-1 accuracy of **79.5%** (31/39), significantly higher than the overall average. Furthermore, the Top-5 accuracy for this group reached **97.4%**, indicating that for these "clear-cut" cases, the correct code is almost always within the immediate candidates retrieved by the lightweight Bi-Encoder. This confirms that the similarity threshold effectively isolates samples where vector representation alone is sufficient for precise mapping.

2) Role of the Fallback Mechanism: Meanwhile, the fallback mechanism successfully captured the remaining **86.3%** of samples (246/285). It is important to note that the Top-1 accuracy for this group was 56.1%, which is lower than the high-confidence group. This disparity is expected and desirable; it indicates that the routing mechanism correctly filtered "hard" or "ambiguous" samples to the more computationally intensive Cross-Encoder. Without this selective routing, these challenging cases would likely have been misclassified by a single-stage retrieval. The robust Top-10 accuracy of 85.3% in this group further suggests that while the exact code is harder to pinpoint, the Cross-Encoder successfully maintains the correct answer within the candidate pool.

C. Sensitivity Analysis of the High-Confidence Threshold

The selection of the confidence threshold τ_h is a critical parameter governing the trade-off between efficiency and accuracy. To determine the optimal value for the high-confidence threshold τ_h, a preliminary validation was conducted using a separate, independent dataset consisting of 36 clinical narratives. By evaluating the trade-off between classification accuracy and routing efficiency on this validation set, $\tau_h = 0.7$ was identified as the optimal operating point and was subsequently fixed for all evaluations on the primary test set. Figure 1 illustrates the system's performance dynamics as τ_h increases from 0.5 to 0.9.

Overall, the retrieval performance exhibits a consistent trend across different Top-k metrics. As τ_h increases, Top-1 accuracy improves significantly, reaching its peak of **59.3%** at $\tau_h = 0.7$. Beyond this point, a gradual decline is observed as the threshold continues to increase. In contrast, Top-5 and Top-10 accuracies show a more stable pattern; they rise initially and plateau around $\tau_h = 0.75$, maintaining consistently high performance without significant degradation at higher thresholds.

From a practical standpoint, the choice of $\tau_h = 0.7$ strikes a balance between computational efficiency and top-ranked accuracy. A higher threshold would route more samples to the computationally intensive Cross-Encoder, increasing inference time, while a lower threshold could significantly reduce overall accuracy. Therefore, $\tau_h = 0.7$ represents a principled operating point that maintains high Top-1 accuracy while conserving resources, making the system suitable for real-world hospital deployment.

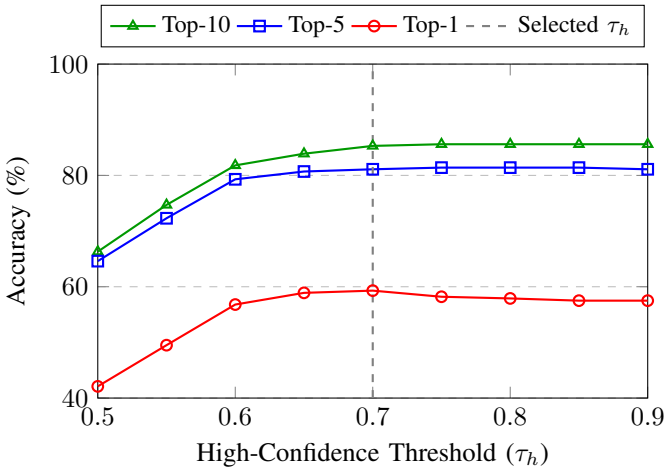

Figure 1. Sensitivity analysis of the high-confidence threshold τ_h. The chart illustrates the variation in Top-1, Top-5, and Top-10 accuracies as the threshold increases from 0.5 to 0.9.

D. Inference Efficiency Analysis

To evaluate the computational efficiency of the proposed hierarchical routing approach, the average inference time per sample for each stage of the system is measured. It should be noted that all experiments were conducted on a personal computer without GPU acceleration (Intel Core i7-10700 CPU, 16 GB RAM), reflecting a resource-constrained clinical environment.

Table III summarizes the runtime comparison among the proposed method, a vanilla Full Cross-Encoder (applied directly to all dictionary candidates), and the baseline BioBERT global retrieval.

TABLE III. Inference time and performance comparison between the baseline (BioBERT global retrieval), the full Cross-Encoder applied to all AIS candidates, and the proposed method. Top-k denotes retrieval accuracy on the full test set.

Method	Top-1 (%)	Top-5 (%)	Top-10 (%)	Avg. Time (ms)
Baseline	39.3	71.6	81.1	**43**
Full Cross-Encoder	56.5	80.0	**86.0**	7,916
Proposed Method	**59.3**	**81.1**	85.3	547

The results show that directly applying a Cross-Encoder to re-rank all candidates in the AIS dictionary is computationally expensive, with an average inference time of 7,916 ms per sample. In contrast, the proposed hierarchical routing strategy substantially reduces the number of candidates forwarded to the Cross-Encoder, leading to a markedly lower average runtime of 547 ms per sample, while preserving strong retrieval performance. This corresponds to an approximately **14.5**× reduction in inference latency compared to the full Cross-Encoder baseline.

To provide a granular view of these efficiency gains, the latency breakdown between the two internal routing paths is analyzed, as detailed in Table IV. The **High-Confidence path** incurs a negligible computational cost of approximately **200 ms** per sample. In contrast, the **Fallback path** (triggering Cross-Encoder re-ranking on the top-k candidates) requires approximately **600 ms** per sample. Crucially, even this intensive fallback path is order-of-magnitude faster than the exhaustive Full Cross-Encoder, as it processes only a limited candidate subset. By successfully routing a portion of clear-cut cases to the fast path, the system achieves the observed weighted average latency of 547 ms.

TABLE IV. Detailed latency breakdown of the proposed hierarchical routing strategy. The system amortizes the cost of the computationally intensive re-ranking stage by successfully offloading clear-cut cases to the high-speed path.

Routing Path	Traffic (%)	Avg. Time	Throughput
High-Confidence	13.7%	∼200 ms	∼300 sample/min
Fallback	86.3%	∼600 ms	∼100 sample/min
Overall System	**100.0%**	**547 ms**	**∼110 sample/min**

By comparison, the baseline BioBERT global retrieval achieves the fastest inference time at 43 ms per sample; however, this efficiency comes at the cost of inferior Top-k accuracy across all evaluation metrics. Notably, the proposed method achieves the highest Top-1 accuracy (surpassing even the full Cross-Encoder) while maintaining inference time within a sub-second range on CPU-only hardware.

At the observed latency level, the system is capable of processing over 100 samples per minute, supporting practical deployment in batch or near–real-time registry workflows.

V. Discussion

This section provides a detailed discussion on the efficiency, routing mechanisms, and clinical implications of the findings.

A. Balancing Efficiency and Accuracy

A key contribution of this study lies in demonstrating that high retrieval accuracy can be achieved without reliance on GPU acceleration, thereby improving the practical deployability of the proposed system. As shown in Table III, the proposed method achieves a favorable balance between inference efficiency and predictive performance. Although its inference time is moderately higher than that of the baseline BioBERT global retrieval, it remains approximately 14× faster than applying the full Cross-Encoder to all AIS candidates, while simultaneously achieving superior Top-1 accuracy (59.3% vs. 56.5%).

Notably, the proposed method even outperforms the full Cross-Encoder in Top-1 accuracy, a result that may appear counterintuitive given the Cross-Encoder's strong modeling capacity. This improvement can be attributed to the role of the first-stage Bi-Encoder, which effectively acts as a **semantic noise filter**. By eliminating semantically irrelevant or weakly related candidates early in the pipeline, the Bi-Encoder restricts the Cross-Encoder's attention to a smaller subset of genuinely challenging and semantically similar candidates. This targeted re-ranking process reduces the likelihood of spurious high-confidence mismatches that may arise when the Cross-Encoder is applied indiscriminately across the entire candidate space.

From a practical perspective, the resulting inference time of **547 ms per query** demonstrates that the proposed hierarchical retrieval strategy is feasible for real-world clinical or registry-based applications, even under standard CPU-only hardware constraints. This balance between efficiency and accuracy highlights the proposed method's suitability for deployment in resource-limited settings, where computational efficiency is a critical requirement alongside reliable diagnostic coding performance.

Beyond computational latency, the **training-free nature** of the framework confers a critical operational advantage: **"plug-and-play" deployability**. Unlike supervised approaches that necessitate the collection and annotation of site-specific datasets for model fine-tuning, the proposed retrieval-based system relies solely on the universal semantic knowledge of pre-trained models. This characteristic facilitates seamless cross-institutional portability, allowing the system to be directly scaled to different hospitals without retraining. Consequently, it effectively bypasses the data privacy barriers and regulatory constraints that typically hinder the aggregation of large-scale medical training data, making it an ideal solution for federated or isolated clinical environments.

B. Mechanism of Hierarchical Routing and Threshold Sensitivity

The effectiveness of the proposed hierarchical routing strategy is primarily driven by its ability to adaptively allocate computational resources based on prediction confidence. As demonstrated in the analysis of sample distribution and classification performance, even though only a relatively small fraction of samples are classified as high-confidence, these cases can be reliably resolved by the first-stage Bi-Encoder, allowing the system to reserve Cross-Encoder processing for genuinely ambiguous inputs.

The sensitivity analysis in Figure 1 further elucidates the role of the high-confidence threshold τ_h in balancing precision and coverage within the hierarchical routing framework. As τ_h increases from 0.5 to 0.7, a consistent improvement is observed across all Top-k metrics, indicating that stricter confidence filtering effectively suppresses noisy predictions produced by the first-stage Bi-Encoder. Notably, Top-1 accuracy exhibits the most pronounced gain, rising sharply and reaching its peak at $\tau_h = 0.7$, which suggests that confidence-based routing is particularly beneficial for improving the correctness of the top-ranked prediction.

Beyond $\tau_h = 0.7$, however, further increasing the threshold leads to diminishing returns. While Top-5 and Top-10 accuracies remain relatively stable and plateau at higher thresholds, Top-1 accuracy begins to slightly decline. This trend indicates that an overly conservative threshold may prematurely route borderline-but-correct cases to the fallback stage. As a result, the effectiveness of early confident resolution is slightly reduced. Taken together, these results support the choice of $\tau_h = 0.7$ as a well-balanced operating point that maximizes top-ranked accuracy while maintaining stable higher-rank performance. This selection represents a principled trade-off rather than an arbitrary choice, ensuring that the Cross-Encoder is invoked only when strictly necessary. Such adaptive behavior highlights both the interpretability and practical efficiency of the proposed hierarchical framework.

C. Region-Specific Performance and Error Analysis

The comparative analysis across AIS body regions reveals that the proposed hierarchical routing strategy consistently outperforms the baseline BioBERT global retrieval in most anatomical regions, as shown in Table I. Notably, substantial gains are observed in regions such as the Neck, Thorax, Abdomen, and Extremities, where Top-1 accuracy improves by over 20 percentage points in some cases. These improvements reflect the system's ability to adaptively focus computational resources on ambiguous cases, effectively leveraging the strengths of both the Bi-Encoder and the Cross-Encoder to resolve complex medical semantics.

However, the analysis also highlights certain regions, such as the Cervical and Thoracic Spine, where the baseline method occasionally matches or slightly exceeds the proposed approach. This phenomenon is driven by two critical factors: the **context gap** in short texts and the **lexical mismatch** within the ontology.

First, spinal injury descriptions in clinical notes are often "telegraphic" and extremely concise (e.g., "C2 fx"), lacking the surrounding context required for BERT-based models to derive robust semantic embeddings. In such data-sparse scenarios, the vector representations of anatomically proximate structures (e.g., 'C7' vs. 'T1') become indistinguishable ("vector crowding"), causing the first-stage Bi-Encoder to assign low confidence scores.

Second, a detailed inspection of the AIS 2015 dictionary reveals a significant **ontological coverage gap**. While clinicians rely heavily on alphanumeric shorthands (C1–C7), the presented analysis indicates that among specific spinal segments, **only "C4" is explicitly mentioned** in the dictionary's descriptive text. Other segments are described using formal anatomical nomenclature (e.g., "Atlas", "Axis", "Odontoid"). As noted by Yuan et al. [23], standard pre-trained models struggle to map these informal clinical abbreviations to formal ontology concepts without explicit knowledge injection. Consequently, when routed to the fallback stage, the Cross-Encoder—lacking the necessary synonym mappings—may paradoxically introduce re-ranking errors on these specific borderline cases.

These observations underscore the necessity of targeted improvements for specific anatomical sub-domains. Future iterations could address the identified lexical gaps by integrating advanced **Named Entity Recognition (NER)** techniques. As comprehensively reviewed in [24], deep learning-based NER models excel at extracting and categorizing context-dependent clinical entities. Integrating such an entity-aware pre-processing module would allow the system to explicitly detect informal shorthands (e.g., 'C2') and normalize them to their canonical ontology forms (e.g., 'Axis') prior to retrieval.

Overall, these findings demonstrate that the hierarchical framework not only improves overall Top-k retrieval accuracy but also provides interpretable insights into region-specific performance, enabling targeted optimization for anatomically challenging categories.

D. Clinical Implications and Workflow Integration

The proposed hierarchical routing framework is designed primarily as a decision-support tool rather than a fully automated coding system. Even in high-confidence cases, the system does not directly assign AIS codes; instead, it aims to alleviate the overall workload by enabling a **risk-stratified workflow**.

Visualizing Confidence for Workflow Optimization: In practice, a user interface is envisioned that dynamically adjusts its presentation based on the model's confidence scores.

- **Streamlined Validation (High-Confidence):** For samples routed via the high-confidence path (Score $\geq \tau_h$), the interface displays the Top-10 candidates with a distinct **"High Confidence" indicator**. Since the correct code is highly likely to be ranked at the very top, registrars can operate in a rapid "verification mode," focusing their attention on the first few options to quickly validate and commit the code, thereby accelerating the throughput of routine cases.
- **Assisted Investigation (Fallback):** Conversely, for ambiguous cases handled by the fallback mechanism, the system triggers an **"Ambiguity Alert"** while still presenting the Top-10 candidates. This visual cue explicitly signals the registrar to shift into an "investigation mode." In this state, the Top-10 list serves as a semantic anchor, but the interface encourages the user to utilize auxiliary tools—such as keyword search within the full AIS dictionary or cross-referencing specific EHR notes—to locate the precise code if it is not immediately apparent in the candidate list.

This confidence-driven guidance allows trauma registrars to allocate their primary cognitive resources to the complex minority of cases that truly require human expertise, rather than expending equal effort on every chart. By reducing repetitive lookups for routine injuries, the system supports a semi-automated, human-in-the-loop workflow that balances high-throughput efficiency with rigorous expert oversight.

E. Limitations

Despite the promising results, several limitations should be noted. First, regarding input robustness, while the underlying BioBERT model utilizes sub-word tokenization to handle minor lexical variations, the system's resilience to severe typographical errors or non-standard abbreviations—common in clinical settings—was not empirically evaluated. Such input noise could potentially affect retrieval accuracy, and future work could incorporate dedicated spelling-correction modules or noise-robust training strategies to enhance real-world applicability.

Second, the current implementation assumes a single-label mapping, meaning each diagnostic input corresponds to only one AIS code. Consequently, the system cannot handle compound narratives where multiple injuries are described within a single sentence. Extending the architecture to support multi-label classification or sentence segmentation will be necessary for more complex trauma cases.

Third, a performance disparity exists across anatomical regions. As highlighted in the results, certain areas—particularly the Spine—exhibit lower Top-k accuracy compared to regions such as the Abdomen. This suggests that a uniform retrieval strategy may not be optimal for all body parts, pointing to the potential benefit of region-specific thresholds or specialized pre-processing to better distinguish anatomically similar structures.

Overall, these limitations indicate directions for future improvement, while the current framework still provides substantial utility as a support tool for semi-automated, human-in-the-loop coding workflows.

VI. Conclusion and Future Work

In this study, a hierarchical semantic routing framework for Abbreviated Injury Scale (AIS) coding that combines Bi-Encoder and Cross-Encoder architectures to balance accuracy and computational efficiency is presented. By adaptively allocating processing based on prediction confidence, the system reliably resolves high-confidence cases with the lightweight Bi-Encoder while reserving the more computationally intensive Cross-Encoder for ambiguous inputs. Experimental results on real-world emergency department narratives demonstrate that the proposed approach consistently outperforms the baseline BioBERT global retrieval. Notably, it achieves superior Top-1 and Top-5 accuracy compared to the exhaustive full Cross-Encoder baseline while maintaining comparable Top-10 performance, all while offering a $\sim$14$\times$ speedup relative to the latter.

Crucially, the framework operates as a training-free solution, relying solely on pre-trained semantic knowledge without the need for task-specific fine-tuning. This eliminates the dependency on site-specific annotated datasets, facilitating immediate cross-institutional adoption.

This design enables practical deployment on standard CPU hardware and provides high-confidence candidate suggestions (Top-5, Top-10) to assist trauma registrars, thereby reducing repetitive work and cognitive load in a human-in-the-loop workflow.

Despite these promising results, several limitations remain. The current system assumes single-label inputs and may underperform on compound or multi-injury narratives. Performance disparities across anatomical regions—particularly in the Spine—highlight the need for region-specific strategies or enhanced pre-processing. Additionally, robustness to severe typographical errors or non-standard abbreviations has not yet been empirically quantified.

Future work will explore integrating lightweight syntactic segmentation modules or specialized entity-extraction layers

to decompose compound narratives into single-injury descriptions before retrieval. This would allow the system to handle multi-label scenarios while maintaining the computational efficiency that makes the proposed framework suitable for resource-constrained environments.

Further optimization of region-specific thresholds and entity-aware processing could help mitigate performance gaps across anatomical areas. Overall, the proposed hierarchical approach establishes a strong foundation for semi-automated, interpretable, and efficient AIS coding, with clear potential to streamline clinical workflows and support trauma registrars in accurately capturing injury data.

References

[1] C. Lee, P. Lee, C. Han, J. Hsu, and C. Hu, "From text to code: Predicting abbreviated injury scale 2015 from clinical narratives," in *HEALTHINFO 2025, The Tenth International Conference on Informatics and Assistive Technologies for Health-Care, Medical Support and Wellbeing*, IARIA, Oct. 2025, ISBN: 978-1-68558-312-5. [Online]. Available: https://www.thinkmind.org/library/HEALTHINFO/HEALTHINFO_2025/healthinfo_2025_1_40_90067.html.

[2] Association for the Advancement of Automotive Medicine, *Abbreviated Injury Scale 2015 Dictionary*. Chicago, IL: Association for the Advancement of Automotive Medicine, 2015, AIS 2015.

[3] S. Ji et al., "A unified review of deep learning for automated medical coding," *ACM Computing Surveys*, vol. 56, no. 12, pp. 1–41, Oct. 2024, ISSN: 1557-7341. DOI: 10.1145/3664615. [Online]. Available: http://dx.doi.org/10.1145/3664615.

[4] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, and J. Eisenstein, "Explainable prediction of medical codes from clinical text," 2018, accessed: 2026.05.07. DOI: 10.48550/ARXIV.1802.05695. [Online]. Available: https://arxiv.org/abs/1802.05695.

[5] S. Puts, C. M. L. Zegers, A. Dekker, and I. Bermejo, "Developing an ICD-10 coding assistant: Pilot study using RoBERTa and GPT-4 for term extraction and description-based code selection," *JMIR Formative Research*, vol. 9, e60095–e60095, Feb. 2025, ISSN: 2561-326X. DOI: 10.2196/60095. [Online]. Available: http://dx.doi.org/10.2196/60095.

[6] P. Cao, Y. Chen, K. Liu, J. Zhao, S. Liu, and W. Chong, "Hypercore: Hyperbolic and co-graph representation for automatic ICD coding," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2020. DOI: 10.18653/v1/2020.acl-main.282. [Online]. Available: http://dx.doi.org/10.18653/v1/2020.acl-main.282.

[7] T. Vu, D. Q. Nguyen, and A. Nguyen, "A label attention model for ICD coding from clinical text," 2020, accessed: 2026.05.07. DOI: 10.48550/ARXIV.2007.06351. [Online]. Available: https://arxiv.org/abs/2007.06351.

[8] J. Lee et al., "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," 2019, accessed: 2026.05.07. DOI: 10.48550/ARXIV.1901.08746. [Online]. Available: https://arxiv.org/abs/1901.08746.

[9] M. Sung, H. Jeon, J. Lee, and J. Kang, "Biomedical entity representations with synonym marginalization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2020, pp. 3641–3650. DOI: 10.18653/v1/2020.acl-main.335. [Online]. Available: http://dx.doi.org/10.18653/v1/2020.acl-main.335.

[10] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, and N. Collier, "Self-alignment pretraining for biomedical entity representations," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online: Association for Computational Linguistics, Jun. 2021, pp. 4228–4238. [Online]. Available: https://www.aclweb.org/anthology/2021.naacl-main.334.

[11] L. Liu et al., "A survey on medical large language models: Technology, application, trustworthiness, and future directions," 2024, accessed: 2026.05.07. DOI: 10.48550/ARXIV.2406.03712. [Online]. Available: https://arxiv.org/abs/2406.03712.

[12] Z. Boukhers, A. Khan, Q. Ramadan, and C. Yang, "Large language model in medical informatics: Direct classification and enhanced text representations for automatic ICD coding," 2024, accessed: 2026.05.07. DOI: 10.48550/ARXIV.2411.06823. [Online]. Available: https://arxiv.org/abs/2411.06823.

[13] Z. Zhu et al., "Can we trust AI doctors? a survey of medical hallucination in large language and large vision-language models," in *Findings of the Association for Computational Linguistics: ACL 2025*, Association for Computational Linguistics, 2025, pp. 6748–6769. DOI: 10.18653/v1/2025.findings-acl.350. [Online]. Available: http://dx.doi.org/10.18653/v1/2025.findings-acl.350.

[14] Q. Jin et al., "MedCPT: Contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval," 2023, accessed: 2026.05.07. DOI: 10.48550/ARXIV.2307.00589. [Online]. Available: https://arxiv.org/abs/2307.00589.

[15] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019, pp. 3982–3992.

[16] P. Deka, A. Jurek-Loughrey, and D. P, "Evidence extraction to validate medical claims in fake news detection," in *International Conference on Health Information Science*, Springer, 2022, pp. 3–15.

[17] N. Thakur, N. Reimers, J. Daxenberger, and I. Gurevych, "Augmented SBERT: Data augmentation

method for improving bi-encoders for pairwise sentence scoring tasks," 2020, accessed: 2026.05.07. DOI: 10.48550/ARXIV.2010.08240. [Online]. Available: https://arxiv.org/abs/2010.08240.

[18] R. Nogueira and K. Cho, "Passage re-ranking with BERT," *arXiv preprint arXiv:1901.04085*, 2019.

[19] L. Gao, Z. Dai, and J. Callan, "Coil: Revisit exact lexical match in information retrieval with contextualized inverted list," 2021, accessed: 2026.05.07. DOI: 10.48550/ARXIV.2104.07186. [Online]. Available: https://arxiv.org/abs/2104.07186.

[20] N. Reimers and T. Aarsen. "Hugging face model repository: Cross-encoder/ms-marco-minilm-l-6-v2." License: Apache License 2.0. Retrieved: 2026.05.07, Hugging Face. [Online]. Available: https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-6-v2.

[21] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, and N. Collier. "Hugging face model repository: Cambridgeltl/SapBERT-from-PubMedBERT-fulltext." License: Apache License 2.0. Retrieved: 2026.05.07, Hugging Face. [Online]. Available: https://huggingface.co/cambridgeltl/SapBERT-from-PubMedBERT-fulltext.

[22] P. Deka. "Hugging face model repository: Pritamdeka/BioBERT-mnli-snli-scinli-scitail-mednli-stsb." License: Creative Commons Attribution Non Commercial 3.0. Retrieved: 2026.05.07, Hugging Face. [Online]. Available: https://huggingface.co/pritamdeka/BioBERT-mnli-snli-scinli-scitail-mednli-stsb.

[23] Z. Yuan, Z. Zhao, H. Sun, J. Li, F. Wang, and S. Yu, "Coder: Knowledge infused cross-lingual medical term embedding for term normalization," 2020, accessed: 2026.05.07. DOI: 10.48550/ARXIV.2011.02947. [Online]. Available: https://arxiv.org/abs/2011.02947.

[24] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," 2018, accessed: 2026.05.07. DOI: 10.48550/ARXIV.1812.09449. [Online]. Available: https://arxiv.org/abs/1812.09449.