DRAGON: Dual-Encoder Retrieval with Guided Ontology Reasoning for Medical Normalization

Dao Sy Duy Minh[*1] **and Nguyen Lam Phu Quy**[*1] **and Pham Phu Hoa**[*1]

Tran Chi Nguyen[1] **and Huynh Trung Kiet**[1] **and Truong Bao Tran**[2]
[1]University of Science, Vietnam National University Ho Chi Minh City
[2]University of Economics and Law, Vietnam National University Ho Chi Minh City
{23122041,23122048,23122030,23122044,23122039}@student.hcmus.edu.vn
trantb234102e@st.uel.edu.vn

Abstract

Adverse Drug Event (ADE) normalization to standardized medical terminologies such as MedDRA presents significant challenges due to lexical and semantic gaps between colloquial user-generated content and formal medical vocabularies. This paper presents our submission to the ALTA 2025 Shared Task on ADE normalization, evaluated using Accuracy@k metrics. Our approach employs distinct methodologies for the development and test phase. In the development phase, we propose a three-stage neural architecture: (1) bi-encoder training to establish semantic representations, (2) lexical-aware fine-tuning to capture morphological patterns alongside semantic similarity, and (3) cross-encoder re-ranking for fine-grained discrimination, enabling the model to leverage both distributional semantics and lexical cues through explicit interaction modeling. For the test phase, we utilize the trained bi-encoder from stage (1) for efficient candidate retrieval, then adopt an alternative re-ranking pipeline leveraging large language models with tool-augmented retrieval and multi-stage reasoning. Specifically, a capable model performs reasoning-guided candidate selection over the retrieved top-k results, a lightweight model provides iterative feedback based on reasoning traces, and an automated verification module ensures output correctness with self-correction mechanisms. Our system achieves competitive performance on both development and test benchmarks, demonstrating the efficacy of neural retrieval-reranking architectures and the versatility of LLM-augmented neural pipelines for medical entity normalization tasks.

1 Introduction

Analyzing Adverse Drug Events (ADEs) from patient-generated text is crucial for pharmacovigilance, but normalizing informal mentions to standardized terminologies like MedDRA remains a major bottleneck. This task is difficult due to the large lexical and semantic gap between colloquial language and formal clinical terms. Successfully normalizing these mentions is essential for data aggregation, interoperability, and downstream safety analyses.

The entity normalization task exhibits inherent complexity due to the substantial lexical semantic divergence between colloquial expressions in patient narratives and the formal, clinically precise terminology in medical ontologies. Users may describe adverse events using varied linguistic realizations, ranging from symptom-focused descriptions ("my stomach hurts badly") to outcome-oriented expressions ("ended up in ER") — that must be mapped to canonical concept identifiers. Vocabulary mismatch, morphological variations, abbreviations, and the inherent ambiguity of natural language further exacerbate this many-to-one alignment problem.

The ALTA 2025 Shared Task (Mollá et al., 2025) addresses this challenge by providing a benchmark for ADE mention normalization to MedDRA Preferred Terms. Participants are tasked with developing systems that, given user-generated text with pre-identified ADE spans, produce ranked lists of candidate MedDRA concepts. System performance is assessed using Accuracy@k metrics, with Accuracy@1 serving as the primary evaluation criterion, alongside Accuracy@5 and Accuracy@10 as auxiliary measures.

Our submission explores two complementary paradigms for medical entity normalization. For the development phase, we develop a three-stage cascaded neural pipeline. The first stage employs a dual-encoder built upon SapBERT-from-PubMedBERT, which projects both ADE mentions and MedDRA terminologies into a shared semantic space via momentum contrastive optimization with strategic hard negative sampling. In the second stage, we conduct lexical-aware refinement

[*]Equal contribution

by leveraging BM25 and TF-IDF-based negative mining, forcing the encoder to capture character-level patterns and surface forms in addition to its semantic understanding. This addresses the challenge of exact string matching and orthographic variations that purely neural approaches often struggle with. The third stage deploys a cross-encoder that performs joint contextualization of mention-concept pairs through bidirectional attention, facilitating direct token interactions for nuanced scoring. During inference, the dual-encoder rapidly screens the entire MedDRA vocabulary for top-k candidates, which the cross-encoder then meticulously re-scores within a manageable pool, achieving an optimal trade-off between retrieval speed and ranking accuracy.

For the test phase, we explore a complementary strategy that synergizes our trained bi-encoder with generative model-based re-assessment. The bi-encoder from stage (1) first rapidly extracts a preliminary candidate set from the complete MedDRA vocabulary. We then deploy a multi-agent system to re-prioritize these candidates through externally-enhanced inference: specifically, Gemini 2.5 Pro Thinking, augmented with web search capabilities, scrutinizes the extracted candidates while articulating transparent logical derivations, after which Gemini 2.5 Flash conducts progressive enhancement by analyzing the inferential pathways. A deterministic quality assurance layer enforces schema adherence and implements autonomous rectification procedures to resolve structural inconsistencies. This integrative approach underscores the synergy achieved by fusing neural candidate extraction with the deliberative inference capacities of contemporary generative models.

Our main contributions are:

- A staged dual-encoder optimization strategy that incorporates surface-form sensitivity into dense semantic representations through deliberate contrastive sampling.
- A robust pairwise scoring architecture for modeling nuanced query-candidate relationships via mutual contextualization mechanisms.
- An innovative composite system unifying rapid neural screening with knowledge-enhanced generative re-assessment via Gemini 2.5 Pro Thinking and Gemini 2.5 Flash, featuring cascaded inference, recursive critique, and programmatic quality control
- Extensive experimental evaluation yielding strong results on the ALTA 2025 benchmark across both traditional learned architectures and integrated neural-generative frameworks.

2 Related Work

2.1 Biomedical Entity Normalization

Biomedical entity normalization maps free-text mentions to standardized ontologies like MedDRA (Sung et al., 2020). While traditional lexical methods are efficient, they fail to bridge the semantic gap between informal patient language and formal medical terms. This has led to a shift towards neural approaches that learn dense semantic representations for more robust matching.

2.2 Dense Retrieval with Contrastive Learning

SapBERT and Biomedical Language Models: (Liu et al., 2021) introduced SapBERT (Self-Alignment Pretraining for BERT), a pivotal biomedical language model that employs metric learning to create semantically meaningful representations. SapBERT leverages synonym relations from ontologies like UMLS to train encoders that position semantically equivalent terms closer in embedding space. This self-alignment pretraining strategy has established SapBERT as a foundational backbone for biomedical entity normalization tasks.

2.3 Bi-encoder and Cross-encoder Architectures

The canonical bi-encoder and cross-encoder framework (Wu et al., 2020) balances efficiency with accuracy. A bi-encoder first retrieves candidates efficiently from a large knowledge base. A more computationally intensive cross-encoder then precisely re-ranks these candidates by modeling direct mention-entity interactions.

This two-stage retrieve-then-rerank paradigm has become the standard approach in entity linking systems, balancing computational efficiency with ranking accuracy. Recent work has extended this framework to biomedical domains, incorporating domain-specific pretraining and specialized negative sampling strategies (Li and Yuan, 2022; Sung et al., 2020).

2.4 Large Language Models for Entity Normalization

LLM-Augmented Normalization and Knowledge-Rich Reasoning: Building on (Dobbins, 2024) multi-stage pipelines, where LLMs generate paraphrases and prune candidates to boost biomedical concept normalization accuracy, and on retrieval-augmented methods that fuse parametric model knowledge with external evidence (Lewis et al., 2021), we synthesize these directions via a hybrid design: efficient bi-encoder retrieval provides a strong candidate set, while LLM-based reasoning adds interpretable, evidence-grounded discrimination that surpasses purely neural or purely generative approaches.

3 Problem definition and Dataset

3.1 Task Formulation

Given a patient-generated narrative document D containing informal descriptions of adverse drug events, and a set of pre-identified mention spans $M = \{m_1, m_2, \ldots, m_n\}$ where each m_i represents a text segment describing a potential ADE, the objective is to map each mention m_i to a ranked list of MedDRA Preferred Terms from a standardized concept vocabulary $\mathcal{C} = \{c_1, c_2, \ldots, c_k\}$.

Formally, for each mention m_i with character offsets $[s_i, e_i]$ in document D, the system must produce a ranked prediction list $\hat{L}_i = [\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_K]$ where $\hat{c}_j \in \mathcal{C}$ and concepts are ordered by decreasing confidence. The ground truth annotation provides a single canonical concept $c_i^* \in \mathcal{C}$ for each mention. System performance is evaluated using Accuracy@k, defined as:

$$\text{Acc@k} = \frac{1}{|M|} \sum_{i=1}^{|M|} \mathbb{1}\left[c_i^* \in \text{top-}k(\hat{L}_i)\right] \quad (1)$$

where $\mathbb{1}[\cdot]$ is the indicator function. The shared task employs Accuracy@1 as the primary metric, with Accuracy@5 and Accuracy@10 serving as secondary evaluation measures.

3.2 Dataset Description

The ALTA 2025 Shared Task dataset comprises patient-authored narratives extracted from online drug review forums, representing authentic real-world adverse event descriptions. The corpus exhibits significant linguistic diversity, encompassing colloquialisms, grammatical inconsistencies, subjective sentiment expressions, and domain-specific abbreviations characteristic of user-generated medical content.

Data Statistics: The dataset is partitioned into training, development, and test splits. The training set contains labeled instances pairing informal ADE mentions with their corresponding MedDRA concept identifiers, enabling supervised model development. The development set facilitates hyperparameter tuning and model selection, while the test set evaluates final system performance. Each instance comprises:

- **Document-level context**: Complete patient narrative providing situational context for adverse event interpretation
- **Mention-level annotations**: Character-offset spans identifying specific ADE descriptions within the narrative
- **Concept mappings**: Ground truth MedDRA Preferred Term identifiers (concept IDs) representing standardized medical terminology

MedDRA Vocabulary: The target concept space consists of the Medical Dictionary for Regulatory Activities (MedDRA) Preferred Terms, a comprehensive hierarchical medical terminology system widely adopted in pharmacovigilance. The vocabulary encompasses thousands of standardized clinical concepts, each uniquely identified by a numerical concept ID paired with a canonical term string (e.g., "10018836": "Haematochezia").

4 Methodology

4.1 Dual-Encoder

The dual-encoder serves as the foundation of our retrieval system, independently encoding ADE mentions and MedDRA concepts into a shared dense representation space. We employ a two-stage progressive training strategy: initial semantic-aware training establishes fundamental conceptual correspondences, followed by lexical-aware refinement that explicitly incorporates surface-level matching signals.

4.1.1 Stage 1: Semantic-Aware Training with Momentum Contrastive Learning

In the initial training stage, we adopt a momentum-based contrastive learning (van den Oord et al.,

2019; Chen et al., 2020; Gao et al., 2022) framework to learn robust semantic representations. The dual-encoder comprises two components: a query encoder $f_q(\cdot)$ for mention embeddings and a key encoder $f_k(\cdot)$ for concept embeddings, both instantiated from a pre-trained biomedical language model backbone.

Momentum Encoder Mechanism: (He et al., 2020) Following the momentum contrast framework, we maintain a momentum-updated key encoder $f_k^{mom}(\cdot)$ that evolves as an exponential moving average of the query encoder parameters:

$$\theta_k^{mom} \leftarrow m \cdot \theta_k^{mom} + (1 - m) \cdot \theta_q \qquad (2)$$

where $m \in [0, 1)$ is the momentum coefficient and θ_q, θ_k^{mom} denote the parameters of query and momentum encoders respectively. This momentum mechanism provides stable and consistent concept representations throughout training, mitigating representation drift as the model parameters evolve.

Contrastive Learning with Hard Negatives: For each training instance consisting of a mention m and its ground-truth concept c^+, we construct a contrastive batch containing one positive pair and multiple hard negative concepts. We optimize using the InfoNCE loss, which maximizes agreement between mention embedding $\mathbf{q} = f_q(m)$ and positive concept embedding $\mathbf{k}^+ = f_k^{mom}(c^+)$ while minimizing similarity to K negative concept embeddings $\{\mathbf{k}_i^-\}_{i=1}^K$. The loss pulls positive pairs closer in the embedding space while pushing apart negative pairs, with similarity measured by cosine distance and controlled by temperature hyperparameter τ.

Hard Negative Mining Strategy: (Xiong et al., 2020) We employ a dynamic hard negative mining procedure to select challenging contrastive examples that accelerate convergence and improve discriminative capacity. At each training epoch, we maintain a fixed-size negative queue $\mathcal{Q}$ populated with concept embeddings from previous batches, providing a diverse pool of hard negatives beyond the current batch. Additionally, we periodically update a global concept embedding cache using the current momentum encoder, enabling efficient retrieval of the most confusable concepts based on semantic similarity to the query mention.

4.1.2 Stage 2: Lexical-Aware Fine-tuning

While the semantic-aware training stage captures high-level conceptual similarities, it may overlook surface-level lexical correspondences crucial for handling exact string matches, abbreviations, and morphological variations. This phenomenon has been observed in neural entity linking and relation extraction systems, where models trained primarily on contextual semantics tend to underweight exact name matching signals (Peng et al., 2020). Similarly, dense retrieval models optimized for semantic similarity can exhibit reduced sensitivity to lexical overlap patterns that prove valuable for matching queries with high surface-form correspondence (Ren et al., 2021). We address this limitation through a lexical-aware fine-tuning stage that explicitly incorporates character-level and token-level matching signals, ensuring the model maintains both semantic understanding and lexical sensitivity.

Lexical-Driven Negative Sampling: Unlike Stage 1's semantic-based hard negative mining, we construct training batches using exclusively lexical retrieval methods. Specifically, for each mention, we retrieve hard negative candidates through:

- **BM25 Retrieval** (Robertson et al., 1995; Robertson and Zaragoza, 2009): A probabilistic term-weighting scheme that scores candidates based on term frequency and inverse document frequency statistics, capturing lexical overlap patterns
- **TF-IDF Retrieval** (Spärck Jones, 1972; Salton and Buckley, 1988) : A classical information retrieval approach emphasizing distinctive terms while penalizing common vocabulary, complementing BM25's scoring mechanism

These lexical methods retrieve concepts sharing surface-level characteristics with the query mention but potentially diverging semantically—precisely the challenging cases where pure neural models struggle. By forcing the model to distinguish between lexically similar but semantically distinct concepts, we compel it to internalize both distributional semantics and explicit string matching patterns.

Continued Training with InfoNCE: We continue optimizing the dual-encoder using the InfoNCE objective , but with negative examples sourced exclusively from lexical retrieval. The training data comprises the same mention-concept pairs as Stage 1, but the negative sampling distribution shifts from semantic similarity to lexical overlap. This curriculum-style progression—from se-

mantic foundations to lexical refinement—enables the model to integrate complementary matching signals without catastrophic forgetting of semantic knowledge.

Negative Pool Refreshing: To maintain training diversity and prevent overfitting to static negative sets, we periodically refresh the lexical negative pool throughout training. At regular intervals, we re-run BM25 and TF-IDF retrieval for all training mentions, incorporating the model's evolving understanding of concept relationships. This dynamic negative sampling ensures the model continuously encounters challenging examples as its discriminative capacity improves.

The resulting dual-encoder, after both training stages, embeds mentions and concepts into a unified space where geometric proximity reflects both semantic relatedness and lexical affinity, enabling robust retrieval across diverse linguistic realizations of adverse events.

4.2 Cross-Encoder Reranking

While the dual-encoder efficiently retrieves candidates through independent encoding, it lacks the capacity to model fine-grained interactions between mention and concept representations. We adopt a cross-encoder reranking architecture following the bi-encoder and cross-encoder framework proposed by (Wu et al., 2020), which has demonstrated strong performance in entity linking tasks.

4.2.1 Cross-Attention Scoring

The cross-encoder processes mention-concept pairs jointly through a single transformer encoder. Given a mention m with surrounding context and candidate concept c, we construct the concatenated input sequence $[\text{CLS}] \oplus m_{\text{ctx}} \oplus [\text{SEP}] \oplus c_{\text{def}} \oplus [\text{SEP}]$, where m_{ctx} incorporates contextual window around the mention span and c_{def} is the concept definition. The transformer's bidirectional self-attention enables explicit token-level interactions, with the final $[\text{CLS}]$ representation projected to a scalar matching score:

$$\text{score}(m, c) = \text{MLP}(\mathbf{h}_{\text{CLS}}) \tag{3}$$

4.2.2 Training with Hard Negative Mining

The cross-encoder is trained using contrastive learning with hard negatives mined from both semantic and lexical retrieval systems. For each training mention m with ground-truth concept c^+, we construct a candidate set by combining:

- **Dual-encoder retrievals**: Top-K_{DE} candidates from the trained bi-encoder, capturing semantically and lexically similar concepts
- **BM25 retrievals**: Top-K_{BM25} candidates from lexical matching, emphasizing surface-level term overlap

This hybrid mining strategy ensures diverse challenging negatives. We sample N hard negatives per positive example and optimize using the InfoNCE loss.

4.2.3 Inference and Reranking

During inference, the dual-encoder first retrieves top-K candidates from the entire MedDRA vocabulary. The cross-encoder then exhaustively scores all K mention-concept pairs through joint encoding, producing refined rankings. This cascaded architecture balances computational efficiency with ranking precision, leveraging the complementary strengths of broad retrieval and fine-grained interaction modeling.

4.3 LLM-based Reranking System

For the test phase evaluation, we explore an alternative paradigm that integrates the trained bi-encoder with large language model-based reasoning for candidate reranking. This approach leverages the generative and reasoning capabilities of contemporary foundation models to perform nuanced semantic matching beyond conventional learned retrieval systems.

4.3.1 Hybrid Retrieval-Reasoning Pipeline

The system operates in a cascaded fashion, combining efficient neural retrieval with deliberative reasoning-based reranking. Given a test mention m, we first employ the bi-encoder from Stage 1 (Section 3.1.1) to rapidly extract a preliminary candidate set $\mathcal{C}_{\text{top-k}} = \{c_1, c_2, \ldots, c_k\}$ from the complete MedDRA vocabulary through dense similarity search. This retrieval phase narrows the search space from thousands of concepts to a tractable subset of candidates requiring fine-grained assessment.

Subsequently, the retrieved candidates undergo iterative reranking through a multi-stage reasoning system. Unlike conventional reranking models that rely solely on learned similarity functions, this pipeline explicitly articulates logical derivations and medical domain reasoning to justify candidate selections, enabling interpretable and evidence-grounded predictions.

4.3.2 Reasoning-Guided Candidate Selection

The primary reasoning component processes each mention-candidate pair through structured analytical reasoning augmented with external knowledge retrieval. Specifically, we employ Gemini 2.5 Pro Thinking (Comanici et al., 2025), a reasoning-optimized language model equipped with web search capabilities, to evaluate the semantic correspondence between informal ADE descriptions and standardized medical terminology.

For each candidate concept $c_i \in \mathcal{C}_{\text{top-k}}$, the model constructs a detailed assessment that includes:

- **Semantic alignment analysis**: Evaluation of conceptual overlap between the colloquial mention and clinical definition
- **External evidence retrieval**: Query-driven web search to gather medical literature, clinical resources, and pharmacological references supporting or refuting the candidate mapping
- **Explicit reasoning chains**: Step-by-step logical derivations articulating why a candidate may or may not represent the correct normalization
- **Confidence scoring**: Probabilistic assessment of mapping correctness based on accumulated evidence

This reasoning process generates transparent justifications for each candidate, facilitating interpretability and enabling downstream refinement based on the logical reasoning traces.

4.3.3 Iterative Refinement through Feedback

To enhance prediction robustness, we introduce a secondary refinement stage that critically analyzes the initial reasoning outputs. We employ Gemini 2.5 Flash (Comanici et al., 2025), a computationally efficient variant optimized for rapid inference, to examine the reasoning traces produced in the previous stage and propose adjustments.

The refinement model receives as input:

- The original mention and surrounding context
- The top-ranked candidates from the reasoning stage
- The explicit reasoning chains justifying each candidate
- The provisional confidence scores

By analyzing these inferential pathways, the refinement model identifies potential logical inconsistencies, overlooked semantic nuances, or insufficient evidence chains. It may adjust candidate rankings, promote undervalued alternatives, or reinforce high-confidence predictions through additional supporting rationales. This iterative critique mechanism serves as a form of self-verification, improving prediction accuracy through multi-perspective evaluation.

4.3.4 Automated Verification and Correction

The final stage implements a deterministic quality assurance layer that ensures structural correctness and format compliance of the system outputs. This verification module performs the following checks:

- **Schema validation**: Ensures output conforms to the required JSON structure with proper mention identifiers and ranked concept lists
- **Concept ID verification**: Validates that all predicted concept identifiers exist in the MedDRA vocabulary
- **Ranking consistency**: Confirms candidates are properly ordered and free of duplicates
- **Completeness checking**: Verifies that predictions exist for all test mentions

When discrepancies are detected—such as malformed concept IDs, invalid rankings, or missing predictions—the module invokes autonomous rectification procedures. These may include programmatic corrections (e.g., removing duplicates, reformatting identifiers) or, for substantive errors, triggering a lightweight reprocessing of the problematic instances through the refinement stage. This quality gate ensures that all submitted predictions meet task specifications while maintaining prediction integrity.

4.3.5 System Integration and Inference

The complete LLM-based pipeline integrates these components into a cohesive reranking system. The inference workflow proceeds as:

1. Bi-encoder retrieves top-k candidates (typically $k = 30 - 50$)
2. Reasoning model evaluates each candidate with external knowledge augmentation and generates justifications

3. Refinement model analyzes reasoning traces and adjusts rankings

4. Verification module validates outputs and applies corrections

5. Final ranked predictions are produced for evaluation

This architecture represents a departure from purely learned retrieval-reranking systems, incorporating symbolic reasoning, external knowledge access, and explicit verification into the entity normalization pipeline. While computationally more expensive than neural-only approaches, the system demonstrates the potential of foundation models with reasoning capabilities for complex semantic matching tasks in specialized domains.

5 Experimental Setup

5.1 Dataset and Preprocessing

We use the ALTA 2025 Shared Task corpus (train/dev/test) with pre-identified ADE spans normalized to MedDRA Preferred Terms. Text is lowercased, punctuation preserved, and mentions are marked in-context with special tags `[MENTION]...[/MENTION]`. For concept side, we index MedDRA PT names plus synonyms/definitions when available. We apply Unicode normalization (NFKC) and strip diacritics for robust matching.

5.2 Baselines

We report a lexical baseline and a bi-encoder and cross-encoder system. Our full development pipeline adds a cross-encoder reranker; the test-phase pipeline replaces the cross-encoder with a multi-agent LLM reranker .

5.3 Development Phase Configuration

We train a SapBERT-based bi-encoder with a two-stage curriculum: (i) semantic pretraining with momentum contrastive learning and ANN-mined hard negatives; (ii) lexical-aware fine-tuning using BM25/TF-IDF negatives. Retrieval uses cosine similarity over mean-pooled embeddings and FAISS for ANN search. A cross-encoder (same backbone) reranks the top-K candidates per mention with pairwise joint encoding.

5.4 Test Phase Configuration

At test time, the trained bi-encoder retrieves top-k candidates. A lightweight multi-agent LLM pipeline performs reasoning-guided reranking with explicit justifications and a deterministic verifier for schema/ID validity. Prompts and guardrails are in Appx. A.

Key Hyperparameters (Main Paper Summary)	
Backbone	`cambridgeltl/SapBERT-from-PubMedBERT-fulltext`
Emb dim / Pooling	768 / mean pooling
Bi-enc Stage 1	15 epochs, batch 32, LR 1×10^{-5}, InfoNCE (τ=0.05), ANN hard negatives (FAISS)
Bi-enc Stage 2	7 epochs, batch 32, LR 2×10^{-5}, InfoNCE (τ=0.07), BM25+TF-IDF negatives
Negatives (Stage 2)	BM25 200 + TF-IDF 200 (pool), 10 negatives/sample, bank 256, remine each epoch
Seq len / Optim	128 tokens / AdamW (wd 0.01, warmup 0.1, grad clip 1.0)
Reranking (Dev)	Cross-encoder on top-K=50 (train top-60; 31 negatives/sample)
Reranking (Test)	Multi-agent LLM pipeline on top-k=30–50 + deterministic verifier

Table 1: Core settings that affect results.

5.5 Evaluation Metrics

We follow the shared task and report Accuracy@1 (primary), Accuracy@5, and Accuracy@10.

6 Results

6.1 Development Phase

We compare three settings on the development split: (i) **Bi-Encoder** (dense retrieval only), (ii) **Cross-Encoder** (our dev-time reranker), and (iii) **LLM (Multi-Agent)** reranking run on the same top-K candidates as the cross-encoder.[1]

Metric	Bi-Encoder (%)	Cross-Encoder (%)	LLM (Multi-Agent) (%)
Accuracy@1	0.2889	0.7975	0.7078
Accuracy@5	0.3996	0.9189	0.8593
Accuracy@10	0.4194	0.9441	0.8985

Table 2: Development split. The LLM multi-agent reranker underperforms the cross-encoder on dev, but is substantially stronger than bi-encoder only. All rerankers consume the same top-K retrieved candidates (here K=50).

6.2 Test Phase

We compare the bi-encoder baseline, cross-encoder reranking, and the final LLM multi-agent pipeline on the shared task test set.

Metric	Bi-Encoder (%)	Cross-Encoder (%)	Multi-Agent Pipeline (%)
Accuracy@1	0.2169	0.2048	**0.3855**
Accuracy@5	0.3855	0.4819	**0.5964**
Accuracy@10	0.4699	0.5181	**0.6506**

Table 3: Test set comparison. Cross-encoder generalizes worse than on development, while the LLM multi-agent pipeline achieves the best Accuracy@1.

[1] For fairness, the LLM reranker on dev does *not* use external web calls; it only reasons over the retrieved candidates and provided definitions/synonyms.

7 Conclusion

Our ALTA 2025 study shows that fusing lexical cues with semantic representations markedly improves medical entity normalization. Lexical-aware fine-tuning boosts recall on surface-overlap mentions—evidence that semantic-only models underweight exact matches—while a cross-encoder re-ranks semantically close candidates and an LLM-augmented stage adds competitive accuracy with interpretable traces for safety-critical review. We reconcile the semantic–lexical tension via staged (curriculum-style) training that progressively encodes both signals; latency remains a practical challenge. Overall, robust normalization requires multi-signal integration: efficient neural screening, deliberative reasoning, and structured knowledge working in concert to bridge patient language and clinical ontologies.

Limitations

Despite achieving strong results, our work has several limitations:

Limited Reranking Evaluation. We have not evaluated the effectiveness of our multi-agent reranking pipeline at larger scales (top-$k > 50$). Understanding how reranking performance scales with candidate set size is important for practical deployment scenarios where initial retrieval may return hundreds of candidates.

Pipeline Generalization Gap. Our bi-encoder + cross-encoder pipeline does not yet generalize consistently across development and test sets, exhibiting performance variance between these splits. This indicates potential overfitting during pipeline optimization or insufficient diversity in training data. Developing architectures that maintain stable performance across different data distributions remains an open challenge requiring further investigation into regularization techniques, data augmentation, and ensemble strategies.

These limitations suggest directions for future work, including curriculum learning for reranking, cross-validation for robust hyperparameter selection, and meta-learning approaches for improved generalization.

Acknowledgments

We thank the Australasian Language Technology Association (ALTA) and the organizers of the ALTA 2025 Shared Task for providing the benchmark dataset and facilitating this research initiative.

We are grateful to the anonymous reviewers for their constructive feedback and insightful suggestions that improved the quality of this work.

We extend our appreciation to the creators and maintainers of the MedDRA terminology system and the developers of open-source libraries including Hugging Face Transformers, FAISS, and the biomedical NLP community for making their pretrained models publicly available. Special thanks to the authors of SapBERT and related biomedical language models whose foundational work enabled our research.

References

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. *Preprint*, arXiv:2002.05709.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Nicholas J Dobbins. 2024. Generalizable and scalable multistage biomedical concept normalization leveraging large language models. *Preprint*, arXiv:2405.15122.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2022. Simcse: Simple contrastive learning of sentence embeddings. *Preprint*, arXiv:2104.08821.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. *Preprint*, arXiv:1911.05722.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Yang Li and Jiawei Yuan. 2022. Generative data augmentation with contrastive learning for zero-shot stance detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6985–6995, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. *Preprint*, arXiv:2010.11784.

Diego Mollá, Xiang Dai, Sarvnaz Karimi, and Cécile Paris. 2025. Overview of the 2025 ALTA shared task: Normalise adverse drug events. In *Proceedings of the 2025 Australasian Language Technology Association Workshop (ALTA 2025)*, Sydney, Australia.

Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2020. Learning from Context or Names? An Empirical Study on Neural Relation Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3661–3672, Online. Association for Computational Linguistics.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. RocketQAv2: A joint training method for dense passage retrieval and passage re-ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2825–2835, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Stephen E. Robertson, Susan Walker, Karen Spärck Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1995. Okapi at trec-3. In *Proceedings of the Third Text REtrieval Conference (TREC-3)*, pages 109–126. NIST.

Gerard Salton and Chris Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical entity representations with synonym marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3641–3650, Online. Association for Computational Linguistics.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. *Preprint*, arXiv:1807.03748.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *Preprint*, arXiv:2007.00808.

A Prompts for LLM-based Reranking (ALTA 2025-Compatible)

A.1 Primary Reasoner (Ranking & Justification)

Goal: Rank top-k MedDRA candidates; return `ranked` + `preds`.

```
System:
You are a biomedical normalization expert for ALTA 2025. Given an informal
ADE mention and top-k MedDRA candidates (PT_ID + term + short def/synonyms),
produce (i) a reasoned ranking and (ii) the official "preds" list of PT_IDs.

Developer Rules:
- Think step-by-step internally; OUTPUT JSON ONLY.
- Evidence > surface overlap (semantics, site, acuity, drug-causality).
- Penalize near-miss (wrong organ/scope). Prefer correct granularity.
- If uncertain, still rank with lower confidence.
- "preds" is used for scoring; use only {{allowed_ids}}; unique; len <= {{k}}.

User Input:
doc_id: {{doc_id}}
mention_index: {{mention_index}}
mention_text: {{mention_text}}
context_text (optional): {{context_text}}
top_k_candidates (k={{k}}):
{{#each candidates}}
- id: {{this.id}}
  term: {{this.term}}
  def_or_syns: {{this.def_or_syns_1line}}
{{/each}}

Return STRICT JSON:
{
  "ranked": [
    {"id":"PT_ID","term":"PT_TERM","confidence":0.0_to_1.0,
     "rationale":"<=2 sentences, concrete clinical cues"}
  ],
  "preds": ["PT_ID","PT_ID2","... up to k ..."]
}
```

A.2 Critic/Refiner (Logic Fix & Reorder)

Goal: Audit Reasoner; fix scope/site; keep schema.

```
System:
You are a rigorous biomedical reviewer. Improve ordering/confidence while
```

```
keeping the SAME JSON schema ("ranked" + "preds").

Checklist:
- Definition alignment (scope, organ/site).
- Granularity (avoid overly broad PTs for specific mentions).
- Lexical traps (high overlap but wrong concept) -> demote.
- Update rationales when changing order/confidence.
- Restrict to {{allowed_ids}}; "preds" unique; len <= {{k}}.

Inputs:
- Reasoner JSON: {{reasoner_json}}
- doc_id/mention_index: {{doc_id}} / {{mention_index}}
- mention/context: {{mention_text}} / {{context_text}}

Return corrected JSON with keys: ranked, preds.
```

A.3 Deterministic Verifier (Schema & ID Guardrail)

Goal: Enforce submission format; drop invalid/duplicate IDs; finalize preds.

```
System:
Strict compliance checker for ALTA 2025.

Rules:
1) Required keys: ranked (array), preds (array of PT_ID strings).
2) Deduplicate ranked by "id" (keep highest confidence).
3) preds = ordered IDs from ranked (highest->lowest).
4) Remove IDs not in {{allowed_ids}}; truncate to <= {{k}}.
5) If ranked becomes empty, synthesize preds using most plausible fallback from {{fallback_terms}} (string-similarity tie-break).
6) Output JSON only.

Input to validate:
{{candidate_json}}

Return final JSON with ranked and preds.
```

A.4 Tool-Augmented Retrieval Planner (Optional)

Goal: Up to 3 compact queries to disambiguate close PTs.

```
System:
Design up to 3 high-precision queries for authoritative sources (MedDRA, EMA, NIH).

User:
Mention: {{mention_text}}
Ambiguous (id|term): {{ambiguous_subset}}

Return:
{"queries":[{"q":"...","why":"A vs B"},
            {"q":"...","why":"verify site"},
            {"q":"...","why":"confirm definition wording"}]}
```

A.5 Safe Fallback (Low-Evidence Cases)

Goal: Conservative ranking with honest uncertainty.

```
If evidence is conflicting/insufficient:
- Choose least-violating candidate; top confidence <= 0.45.
- Keep allowed_ids; preds length <= {{k}}.
- Output JSON only (ranked + preds) with uncertainty noted.
```

A.6 Submission Line Adapter (Per-Mention)

Goal: Emit the exact ALTA submission object per mention.

```
System:
Convert to the ALTA 2025 submission line.

Inputs:
doc_id: {{doc_id}}
mention_index: {{mention_index}}
verified_json: {{verified_json}}  // includes "preds": ["PT_ID",...]

Return STRICT JSON (single object):
{"id":"{{doc_id}}-{{mention_index}}",
"preds":[{{csv_of_verified_ids}}]}
```