SCaLER@ALTA 2025: Hybrid and Bi-Encoder Approaches for Adverse Drug Event Mention Normalization

Shelke Akshay Babasaheb and **Anand Kumar Madasamy**
Dept. of Information Technology, NITK Surathkal
shelkeakshaybabasaheb.242it030@nitk.edu.in
m_anandkumar@nitk.edu.in

Abstract

This paper describes the system developed by Team Scaler for the ALTA 2025 Shared Task on Adverse Drug Event (ADE) Mention Normalization. The task aims to normalize free-text mentions of adverse events to standardized MedDRA concepts. We present and compare two architectures: (1) a Hybrid Candidate Generation + Neural Reranker approach using a pretrained PubMedBERT model, and (2) a Bi-Encoder model based on SapBERT, fine-tuned to align ADE mentions with MedDRA concepts. The hybrid approach retrieves candidate terms through semantic similarity search and refines the ranking using a neural reranker, while the bi-encoder jointly embeds mentions and concepts into a shared semantic space. On the development set, the hybrid reranker achieves Accuracy@1 = 0.3840, outperforming the bi-encoder (Accuracy@1 = 0.3298). The bi-encoder system was used for official submission and ranked third overall in the competition. Our analysis highlights the complementary strengths of both retrieval-based and embedding-based normalization strategies.

1 Introduction

Adverse Drug Event (ADE) monitoring plays a critical role in pharmacovigilance and post-marketing surveillance (Usui et al., 2018). Detecting and normalizing mentions of ADEs from user-generated text such as social media and medical forums allow healthcare professionals to track drug safety signals in real time. Normalization — mapping a noisy, user-written mention (e.g., “stomach upset”) to a standardized medical concept (e.g., “Nausea” in MedDRA) — remains a challenging task due to synonymy, ambiguity, and lexical variation.

The ALTA 2025 Shared Task (Mollá et al., 2025) focuses specifically on ADE mention normalization to MedDRA terminology. Participants are required to produce a ranked list of possible MedDRA terms for each ADE mention, evaluated using Accuracy@n metrics. To address this challenge, our team developed and evaluated two complementary systems:

- A Hybrid Candidate Generation + Neural Reranker System, built using pretrained PubMedBERT (Gu et al., 2021), which retrieves and refines candidate terms without retraining.
- A Bi-Encoder System based on SapBERT (Liu et al., 2020), which learns joint embeddings for ADE mentions and MedDRA terms via contrastive training.

Although both systems show strong performance, the hybrid reranker provided better precision.

2 Task Description and Dataset

This section details the problem definition, the structure of the dataset used for this task, and the metrics for evaluation.

2.1 Task Definition

The core objective is normalization of Adverse Drug Event (ADE) mentions from unstructured, user-generated text. Entity Normalization means mapping each ADE mention to its corresponding canonical concept(s) in the Medical Dictionary for Regulatory Activities (MedDRA) terminology (e.g., normalizing "extreme muscle pain" to the MedDRA Preferred Term Myalgia (ID 10028411)). The challenge lies in handling the high lexical diversity and colloquial language present in patient reviews and mapping them to a standardized medical vocabulary.

2.2 Dataset Description

The dataset is partitioned into training, development, and test sets, accompanied by a medDRA.json dictionary file. Each data instance in the train, dev, and test splits represents a single document (a patient's review) and is formatted as a JSON object with the following structure:

- doc_id: A unique string identifier for the document.
- text: The full, raw text content of the patient review.
- mentions: A list of annotated ADEs found within the document.Each annotation in this list contains:
  - text: The exact substring from the document corresponding to the ADE mention.
  - offsets: The start and end character indices of the mention, locating it precisely within the parent text.
  - concepts: A dictionary mapping one or more MedDRA Concept IDs to their official Preferred Term names.

A key characteristic of the dataset is that a single document can contain multiple distinct ADE mentions, and a single mention can sometimes be mapped to multiple MedDRA concepts to capture its full semantic meaning. The medDRA.json file serves as the comprehensive knowledge base, containing over 20,000 Preferred Terms that constitute the target vocabulary for the normalization task. This large target space frames the normalization challenge as a large-scale semantic retrieval or classification problem.

2.3 Evaluation Metrics

System performance for the normalization task is measured using top-k accuracy, which evaluates a model's ability to rank the correct concept highly among all possible candidates. The primary metrics are : Accuracy@1 (Acc@1) : The percentage of mentions where the top-ranked prediction is the correct MedDRA concept. Accuracy@5 (Acc@5) : The percentage of mentions where the correct MedDRA concept appears within the top-5 ranked predictions. Accuracy@10 (Acc@10) : The percentage of mentions where the correct MedDRA concept appears within the top-10 ranked predictions. These metrics effectively measure both the precision of the top prediction and the system's broader recall capabilities within a ranked list.

3 Related Work

Early research on medical concept normalization emphasized developing annotated corpora tailored to specific domains. Luo et al. (2019) introduced the MCN corpus for formal clinical text, highlighting challenges such as compositional concepts and hierarchical mappings using terminologies like SNOMED CT. In contrast, Karimi et al. (2015) created the CADEC corpus for patient-generated content, revealing the complexity of informal, noisy language. The domain generalization gap between such datasets was later quantified by Dai et al. (2024) through the MultiADE benchmark, showing that models trained on one domain often fail to generalize to others. Methodologically, normalization has progressed from symbolic systems like MetaMap to neural architectures leveraging representation learning. Zhang et al. (2022) advanced this field with KRISSBERT, a self-supervised contrastive model that learns domain-agnostic biomedical semantics. Recently, Xiao et al. (2023) introduced INSGENEL, an instruction-tuned generative entity linking framework that equips large language models with entity linking capability via a sequence-to-sequence EL (entity linking) objective and a lightweight mention retriever, achieving substantial efficiency gains while mitigating generative hallucinations. Parallelly, bi-encoder architectures such as SapBERT Liu et al. (2020) and PubMedBERT Gu et al. (2021) have emerged as efficient and scalable alternatives, leveraging self-alignment pretraining and domain-specific language modeling respectively to align medical mentions with ontology concepts, thereby enabling rapid and effective semantic retrieval across large biomedical knowledge bases.

4 System Overview

We developed two distinct systems for the Adverse Drug Event (ADE) normalization task. The first is a multi-stage hybrid retrieval and reranking pipeline, while the second is an end-to-end dense retrieval system based on a bi-encoder architecture. Both systems leverage transformer models pre-trained on biomedical corpora.

4.1 System 1: Hybrid Candidate Generation + Neural Reranker

This system follows a two-stage "retrieve-then-rerank" paradigm. It is designed to first cast a wide net to gather a diverse set of potential candidates and then use a powerful, fine-grained model to identify the best match.

4.1.1 Stage 1: Hybrid Candidate Generation

Instead of relying on a single retrieval method, we implemented a sophisticated hybrid strategy that combines three distinct approaches to generate a robust set of initial candidates for each ADE mention:

1. **Lexical Matching (Fuzzy Search):** We employ the `rapidfuzz` library to perform a token-based fuzzy string search (`fuzz.ratio`) between the mention and the entire corpus of MedDRA synonym names. This captures candidates with high surface-level similarity.

2. **Sparse Retrieval (BM25):** A **BM25Okapi** index (Robertson et al., 1994) is built over all MedDRA synonyms. Given a mention, it is tokenized, and candidates are retrieved based on term frequency and inverse document frequency (TF-IDF), a classic and effective keyword-based retrieval method.

3. **Dense Retrieval (Semantic Search):** We use a pre-trained `microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext` model to compute a 768-dimensional embedding for every MedDRA synonym. These embeddings are indexed using a **FAISS `IndexFlatIP`** structure (Johnson et al., 2017) for efficient similarity search. Mention embeddings are computed on-the-fly and compared against the index to find semantically similar concepts.

The results from these three methods are combined using a weighted voting scheme (Fuzzy: 3, BM25: 2, Semantic: 1) to produce a final candidate list of the top 50 concepts. This ensemble approach ensures that we capture candidates that are lexically, statistically, and semantically relevant.

4.1.2 Stage 2: Neural Reranking

The top 50 candidates are then re-ranked using a powerful cross-encoder model. This reranker is also based on `microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext`, but it is fine-tuned on the competition's training data.

- **Input Formulation:** For each `(mention, candidate)` pair, we create a single input sequence for the model formatted as: `[CLS] mention_text[SEP]candidate_concept_name[SEP]`

- **Training:** The model is trained as a binary sequence classifier (`AutoModelForSequenceClassification`) to predict whether a candidate is the correct normalization for a given mention. Training data is constructed using hard negative sampling. For each positive pair, we sample 10 negative examples from the candidates retrieved in Stage 1 that are not the gold-standard concept.

- **Inference:** During inference, each of the 50 candidates is scored by the reranker. The final output is the list of candidates sorted in descending order of their predicted relevance scores.

4.1.3 Post-processing: Heuristic Tie-Breaking

As a final step, we apply a simple but effective heuristic to refine the ranking. We maintain a list of common anatomical keywords (e.g., "foot", "leg", "eye"). If a mention contains one of these keywords, we boost the rank of any candidate concepts whose names also contain that keyword.

4.2 System 2: Bi-Encoder (SapBERT)

Our second system, is a more streamlined dense retrieval approach using a bi-encoder architecture. This system is trained end-to-end to map mentions and MedDRA concepts into a shared, semantically meaningful embedding space.

4.2.1 Model Architecture

We selected `cambridgeltl/SapBERT-from-PubMedBERT-fulltext` as our base model. SapBERT is particularly well-suited for this task as it was pre-trained using a self-alignment objective on biomedical synonyms, making it highly effective at learning similarities between different phrases for the same concept. The model functions as a dual or bi-encoder, generating separate embeddings for the mention and the concept.

4.2.2 Two-Phase Training Procedure

To maximize performance, we implemented a two-phase training strategy:

1. **Phase 1: Initial Training with In-Batch Negatives:** The model is first trained on the provided training set. For each positive `(mention,concept)` pair in a batch, all other concepts within that same batch are treated as negatives. The model is optimized using a **contrastive loss** function (specifically, cross-entropy over the similarity scores), which

pushes the embeddings of positive pairs closer together and pulls negative pairs apart in the vector space.

2. **Phase 2: Hard Negative Mining and Retraining:** After the initial training phase, we use the trained model to "mine" for hard negatives. We process the entire training set again, and for each mention, we retrieve the top-k candidates from our MedDRA knowledge base. We identify cases where the correct concept is retrieved but is not the top-ranked result. These higher-scoring incorrect concepts are collected as explicit hard negatives. The model is then retrained for additional epochs using these curated, difficult examples, further refining its ability to make fine-grained distinctions.

4.2.3 Retrieval and Inference

During inference, the embeddings for all 74,000+ MedDRA concepts are pre-computed and stored in a **FAISS `IndexFlatIP`** index. For each test mention, its embedding is generated and used to query the index via Maximum Inner Product Search (MIPS). This efficiently returns the top-k most similar MedDRA concepts as the final predictions.

5 Experimental Setup

5.1 Implementation Details

The key components and training details for both systems are outlined in Table 1.

Table 1: Comparison of System-Specific Configurations.

Component	**System 1**	**System 2**
Candidate Generator	Hybrid ensemble of BM25, RapidFuzz, and pre-trained PubMedBERT embeddings.	Not applicable (end-to-end Dual Encoder).
Encoder Architecture	Dual Encoder for candidate generation and a Cross-Encoder for reranking.	Dual Encoder architecture.
Base Model	`microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext`	`cambridgeltl/SapBERT-from-PubMedBERT-fulltext`
Training Strategy	The reranker is fine-tuned; the candidate generator relies on pre-trained models without further training.	Fine-tuned in two distinct phases, including a hard-negative mining step.
Loss Function	Binary Cross-Entropy for the reranker's classification task.	A contrastive loss, implemented via Cross-Entropy over similarity scores.

Both systems were implemented using PyTorch and the HuggingFace Transformers library (Wolf et al., 2019). The following configuration details were common to both training pipelines:

- **Optimizer:** AdamW with a learning rate of 2×10^{-5}.
- **Hyperparameters:** All models were trained for **10 epochs** with a batch size of **16**.
- **Hardware:** Training and inference were conducted on a Kaggle notebook equipped with GPU T4 x 2.

5.2 Inference Pipeline

For each mention in the test set, the final predictions are generated as follows:

- **System 1 (Hybrid + Re-Ranker):**
  1. Generate top 50 candidates using the weighted ensemble of BM25, fuzzy search, and FAISS.
  2. Score all 50 candidates with the fine-tuned cross-encoder reranker.
  3. Apply the anatomical keyword tie-breaking heuristic.
  4. Return the final sorted list of top 10 candidates.

- **System 2 (Bi-Encoder):**
  1. Encode the mention (with context) into a SapBERT embedding.
  2. Perform a MIPS query against the pre-computed FAISS index of all MedDRA concepts.
  3. Return the top 10 results directly from the search.

6 Results and Analysis

We evaluated our two systems on the development set using the official top-k accuracy metrics. The performance of each system is summarized in Table.

Table 2: Performance comparison on the dev set.

System	**Acc@1**	**Acc@5**	**Acc@10**
Bi-Encoder	0.3298	0.7338	0.8457
Hybrid+Reranker	0.3840	0.4417	0.4605

6.1 Performance Analysis

The results highlight a clear trade-off between the precision-oriented reranking approach and the recall-oriented retrieval approach.

Our Hybrid + Neural Reranker system achieved the highest Accuracy@1 (0.3840), demonstrating its superior ability to pinpoint the single correct concept from a list of candidates. This strong precision is attributable to the cross-encoder architecture of the reranker, which performs deep, token-level interaction between the mention and each candidate concept. This fine-grained analysis allows the model to better resolve subtle semantic distinctions that a bi-encoder might miss.

Conversely, the Bi-Encoder (SapBERT) system significantly outperformed the hybrid system on Accuracy@5 (0.7338) and Accuracy@10 (0.8457). This indicates that while it may not always place the correct concept at the very top rank, its ability to retrieve a set of highly relevant candidates is exceptionally strong. The end-to-end training on the task, coupled with the SapBERT model's inherent strength in aligning biomedical terms in a shared vector space, results in excellent recall. The system consistently places the correct concept within the top few results, making it highly effective for tasks where a small, high-quality candidate set is sufficient.

6.2 Discussion and Official Submission

Our key observation is that the two architectures exhibit complementary strengths. The hybrid system's success, even with a pre-trained (not fine-tuned) candidate generator, underscores the power of combining diverse retrieval signals (lexical, sparse, and dense). The reranker then acts as a highly effective "judge" for this pre-selected set. The bi-encoder's performance, on the other hand, confirms that end-to-end dense retrieval is a powerful and efficient method for capturing broad semantic relevance.

Given its superior top-10 precision, the Bi-Encoder system was selected for our official submission. This system ultimately achieved third place on the official ALTA 2025 shared task leaderboard, confirming its effectiveness.

7 Conclusion and Future Work

In this work, we presented and evaluated two complementary systems for the task of Adverse Drug Event (ADE) mention normalization. Our first system was a multi-stage Hybrid Candidate Generation and Neural Reranker pipeline based on PubMedBERT. The second was an end-to-end Bi-Encoder model using SapBERT, which was fine-tuned on the task-specific ADE–MedDRA pairs.

Our results demonstrate the effectiveness of the hybrid approach and bi-encoder approach. System 1 system achieved the best overall top-1 accuracy on the development set ($Acc@1 = 0.3840$) whereas system 2 achieved best top-10 accuracy($Acc@1 = 0.8457$) and we secured third place on the official ALTA 2025 shared task leaderboard.

For future work, we have identified several promising directions to build upon our current results:

- Integrate contextual drug information to resolve ambiguous mentions.
- Employ contrastive hard-negative mining and knowledge distillation between bi-encoder and cross-encoder models.
- Explore instruction-tuned medical LLMs for zero-shot normalization.

Acknowledgments

We thank the ALTA 2025 Shared Task organizers for providing the dataset and evaluation infrastructure. We also acknowledge the open-source communities behind HuggingFace Transformers, PyTorch, and FAISS, which enabled efficient experimentation.

References

Xiang Dai, Sarvnaz Karimi, Abeed Sarker, Ben Hachey, and Cecile Paris. 2024. Multiade: A multi-domain benchmark for adverse drug event extraction. *Journal of Biomedical Informatics*, 160:104744.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *CoRR*, abs/1702.08734.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of Biomedical Informatics*, 55:73–81.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2020. Self-alignment pre-training for biomedical entity representations. *CoRR*, abs/2010.11784.

Yen-Fu Luo, Weiyi Sun, and Anna Rumshisky. 2019. Mcn: A comprehensive corpus for medical concept normalization. *Journal of Biomedical Informatics*, 92:103132.

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 Workshop (ALTA 2025)*, Sydney.

Stephen Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at trec-3. pages 0–.

Misa Usui, Eiji Aramaki, Tomohide Iwao, Shoko Wakamiya, Tohru Sakamoto, and Mayumi Mochizuki. 2018. Extraction and standardization of patient complaints from electronic medication histories for pharmacovigilance: natural language processing analysis in japanese. *JMIR medical informatics*, 6(3):e11021.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Zilin Xiao, Ming Gong, Jie Wu, Xingyao Zhang, Linjun Shou, and Daxin Jiang. 2023. Instructed language models with retrievers are powerful entity linkers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2267–2282, Singapore. Association for Computational Linguistics.

Sheng Zhang, Hao Cheng, Shikhar Vashishth, Cliff Wong, Jinfeng Xiao, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Knowledge-rich self-supervision for biomedical entity linking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 868–880, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.