Team MonoLink at the ALTA Shared Task 2025: Synonym-Aware Retrieval with Guideline-Aware Re-Ranking for MedDRA Normalization

James C. Douglas
james123douglas@gmail.com

Abstract

We describe Team MonoLink's system for the ALTA 2025 Shared Task on normalizing patient-authored adverse drug event (ADE) mentions to MedDRA Lowest Level Terms (LLTs). Our pipeline combines recall-oriented, synonym-augmented candidate retrieval with cross-encoder re-ranking and a guideline-aware LLM discriminator. On the official hidden test set, our submission tied for first place, achieving an Accuracy@1 of 39.8%, Accuracy@5 of 78.3%, and Accuracy@10 of 85.5%.

1 Introduction

Adverse drug event (ADE) surveillance is a core pillar of pharmacovigilance, enabling the recognition, evaluation, and mitigation of therapy-associated harms (Beninger, 2018, 2020). The Medical Dictionary for Regulatory Activities (MedDRA) is a standardized, internationally maintained and adopted terminology that supports this process (Brown et al., 1999; Mozzicato, 2009). It organizes adverse events along a five-level hierarchy to facilitate aggregated analyses (Zink et al., 2025). Lowest Level Terms (LLTs) are the most granular level, designed to capture an ADE mention's original phrasing. In the latest MedDRA edition (v28.1), there are 81,143 active LLTs (MedDRA MSSO, 2025).

Beyond formal reporting channels, patient-authored text from discussion forums and social media provides complementary coverage, often capturing events that are underreported in clinical settings (Golder et al., 2015; Pappa and Stergioulas, 2019; Golder et al., 2024). However, mapping these ADE mentions to regulated vocabularies such as MedDRA is nontrivial because of misspellings, figurative language, and nonstandard phrasing (Khan et al., 2025).

The sixteenth ALTA Shared Task (2025) targeted normalizing (linking) patient ADE mentions in English text to MedDRA LLTs (Mollá et al., 2025). Given a forum-style post and one or more pre-extracted ADE mentions, systems were required to return, for each mention, a ranked list of MedDRA terms judged to be the best mappings. Systems were ranked by Accuracy@1, with Accuracy@5 and Accuracy@10 reported as reference metrics. McNemar's test was applied to top-1 outcomes to test for significant differences between submissions.

This paper presents **Team MonoLink's system** for the shared task. Our approach consists of a synonym-aware candidate retrieval phase, followed by a candidate re-ranking phase that attempts to discriminate between closely related MedDRA terms. On the task's hidden test set, our system achieved the highest scores on the official and reference metrics and was a joint winner. Ordering the top MedDRA terms for each ADE mention proved to be a challenging task, with the system producing an Accuracy@1 of 39.8% and an Accuracy@5 of 78.3%.

2 Task Definition and Data

2.1 Dataset and Annotations

The development data were derived from the CADEC corpus of patient forum posts, with a training set of 4,200 mentions and a validation set of 849 mentions labeled with MedDRA terms (Karimi et al., 2015; Dai et al., 2024). The organizers also supplied a JSON file of MedDRA concepts with numeric identifiers and textual descriptions. For the final evaluation, an unseen test set of 83 mentions was released. Each post was accompanied by sentence boundary annotations, and each mention included character offsets.

2.2 Shared Task Rules

Submissions were required to be fully automatic. External resources (e.g., vocabularies and data) were permitted provided that they did not contain or reveal gold labels for the held-out test instances.

3 Data Preparation and Development Splits

3.1 Preprocessing

Label corrections During early error analysis, we identified a small number of clear, repeated annotation errors in the official training/validation splits (e.g., "General weakness in my calves" labeled as *Flatulence*; "mood swings" labeled as *Pain in fingers*). To reduce noise during development, we manually corrected 81 training labels and 18 validation labels.

Spelling correction We applied a lightweight correction pass using `symspellpy` (v6.9.0), with a conservative maximum edit distance of 1 to resolve clear misspellings while minimizing changes to colloquialisms. Before correction, we removed nonalphanumeric characters (e.g., #, !, @) and collapsed repeated letters (e.g., *sooooo* $\rightarrow$ *so*). For the reference dictionary, we combined the package's default English lexicon with medical terms from UMLS 2025AA (MRCONSO), sourcing from MedDRA, SNOMED CT US, CHV, HPO, and MeSH (Bodenreider, 2004; U.S. National Library of Medicine, 2025). Medical term frequencies were up-weighted to prioritize them over general-English suggestions.

Context extraction To address occasional sentence-boundary errors, we re-segmented posts with `PyRuSH` through `MedSpaCy` (v1.3.1). For each ADE mention, we retained its source sentence and a ± 1 sentence window as context for re-ranking.

MedDRA to UMLS alignment We mapped each task-supplied MedDRA concept to a UMLS Concept Unique Identifier (CUI) using UMLS 2025AA (MRCONSO) (Bodenreider, 2004; U.S. National Library of Medicine, 2025). If a MedDRA concept had no CUI entry ($\sim$0.1% of cases), we retained its MedDRA identifier. We then expanded each CUI with additional synonyms from SNOMED CT US and the Consumer Health Vocabulary (CHV). The resulting synonym table contained 127,919 terms covering 48,315 unique concept identifiers ($\sim$2.65 synonyms per concept).

3.2 External Datasets and Augmentation

To expand the pool of data available for development, we incorporated additional public datasets following the same cleaning and UMLS alignment approach:

- **SMM4H-2017**: ADE mentions from social media (tweets) (Sarker et al., 2018).
- **PsyTAR**: patient forum ADE mentions (Zolnoori et al., 2019).
- **MedNorm**: an aggregated dataset for medical concept normalization (Belousov et al., 2019).

To avoid leakage from CADEC data within MedNorm, we only used data from TwADR-L (Limsopatham and Collier, 2016), TwiMed (Alvaro et al., 2017), and TAC 2017 ADR (Demner-Fushman et al., 2018).

3.3 Generalization Set for Model Selection

We identified a notable overlap between the official training and validation splits. Of 288 unique MedDRA concepts in the validation set, 219 (76%) also appeared in training, and 35.3% of mentions matched training mentions verbatim. To more critically assess system generalization during development, we constructed a **generalization set** by sampling mentions that (i) map to less frequent concepts (CUI frequency below the 75th percentile across all datasets) and (ii) favor longer spans by stratifying into length bins. The final set comprised 486 mentions: 231 from the official training split, 51 from the official validation split, and 204 from SMM4H-2017. We used this set for model selection and evaluation (*identifiers available upon request*).

4 Method

4.1 Overview

Figure 1 summarizes our system. After preprocessing and synonym-table curation, we retrieve candidate CUIs, and then re-rank at the MedDRA LLT level.

4.2 Stage 1: Candidate Retrieval

We retrieve at the CUI level using a synonym table built from the task-supplied MedDRA terms and the UMLS. Operating at the CUI level (i) enables synonym expansion to increase recall, (ii) reduces the label space by collapsing surface-form variants (e.g., *hand pain* vs. *pain in hand*), and (iii) provides a bridge for integrating external datasets.

Retrieval uses exact nearest-neighbor search (cosine similarity) via `FAISS` v1.12.0. We combine a sparse lexical retriever (character n-gram TF-IDF)

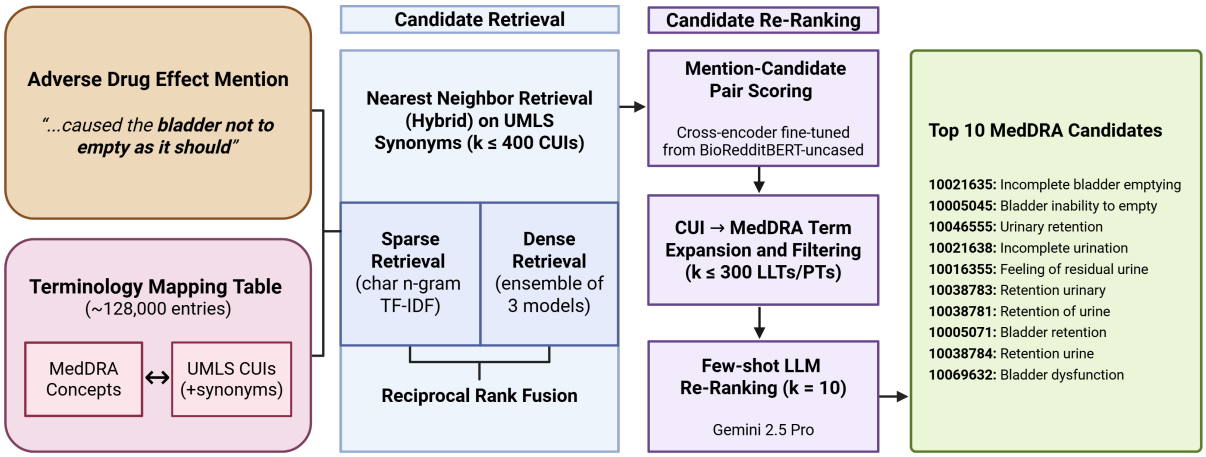

Figure 1: System overview. First-stage retrieval produces up to 400 UMLS CUI candidates per mention. A cross-encoder then re-ranks candidate CUI descriptions, and the highest-scoring subset is expanded to MedDRA LLTs (max. 300). A listwise LLM re-ranker, prompted with MedDRA rules and examples, returns the top 10 LLTs.

with an ensemble of three off-the-shelf dense bi-encoders. Prior work shows that lexical and semantic methods are complementary (Kuzi et al., 2020; Gao et al., 2021).

Character n-gram TF-IDF Mentions and candidate strings are encoded with a character-level TF-IDF model (`scikit-learn` v1.7.1). We use character-level TF-IDF for robustness to misspellings and morphology (preprocessing and hyperparameters in Appendix B). We select the top 100 candidate CUIs per mention.

Dense retrievers We take the union of the top 100 candidates from three off-the-shelf bi-encoders (no fine-tuning):

1. *cambridgeltl/SapBERT-from-PubMedBERT-fulltext* (∼109M parameters), a biomedical embedding model pretrained on UMLS synonyms (Liu et al., 2021).

2. *NovaSearch/stella_en_400M_v5* (∼435M parameters), a general-purpose text embedding model obtained by distillation from larger LLMs (Zhang et al., 2024). We observed best performance using its *s2p* prompt.

3. *ls-da3m0ns/bge_large_medical* (∼335M parameters), a medically adapted variant of the BGE family (Chen et al., 2024).

The models retrieve partially complementary candidates, so we take their union rather than relying on a single model. For a given CUI, we assign it the best rank given by any of the retrievers. We ran retrieval using `Tokenizers` v0.22.0, `Transformers` v4.56.0, `PyTorch` v2.8.0, and `Python` 3.10.18.

Candidate fusion We fuse the sparse and dense ensemble CUI ranks using Reciprocal Rank Fusion (RRF), with a fusion constant of 60. In the process, we deduplicate CUIs but preserve the best-matching dense and sparse synonyms for downstream scoring (these can differ). The theoretical upper bound is 400 unique CUIs per mention (4×100). Because we retain up to two best-matched descriptions per CUI (one from the dense ensemble, one from the sparse retriever), the upper bound on retrieved CUI descriptions is 800. In practice, however, due to overlaps, on the generalization set we observe a median of 250 CUIs per mention (IQR 222–278) and 292 unique description strings (IQR 265–320).

4.3 Stage 2: Expansion and Re-ranking

CUI-level cross-encoder. We fine-tune *cambridgeltl/BioRedditBERT-uncased* as a cross-encoder with `Sentence Transformers` v5.1.0. BioRedditBERT is initialized from BioBERT and further pre-trained on health-related Reddit posts (Basaldella et al., 2020). We train on the candidate retrieval outputs from the official and external datasets, and evaluate using the generalization set (hyperparameters in Appendix C). Training uses a listwise objective (LambdaLoss) with positives from correctly retrieved descriptions or ground-truth CUI synonyms and hard negatives mined from top-ranked false positives. The

positive-to-negative ratio was approximately 1:7.

MedDRA expansion and filtering Cross-encoder scores are used to filter CUIs, balancing recall and candidate count. When two descriptions for a given CUI receive different scores, we keep the highest-scoring term for ranking. CUIs surpassing the threshold are deterministically expanded to their MedDRA term(s) in the order given by the task-provided JSON. We then filter out non-current terms and entries that are neither LLTs nor PTs (using UMLS metadata), and cap the list at the top 300 MedDRA terms per mention to control LLM context.

Few-shot LLM re-ranker We pass the resulting MedDRA candidates for a mention, along with surrounding context, to Gemini 2.5 Pro in a single listwise prompt that returns the indices of the top 10 LLTs. The prompt encodes MedDRA term-selection principles and tie-breaking advice, and includes few-shot examples derived from public training materials and development data. Decoding and prompt details are in Appendix G.

5 Experiments and Results

5.1 Candidate Retrieval

We evaluate first-stage retrieval with Recall@k on the official training/validation splits and on the curated generalization set (Table 1). On the label-corrected training and validation splits, Recall@all is approximately 99%, limiting error propagation to the re-ranking stage. Recall is lower on the generalization set, reflecting the challenge of longer spans and less-frequent concepts.

Split	R@10	R@50	R@100	R@all
Training	81.3	90.8	94.1	97.5
+ label corrections	82.9	92.3	95.6	99.1
Validation	81.0	89.5	92.6	97.4
+ label corrections	83.0	91.6	94.7	99.2
Generalization set	64.4	82.7	87.4	93.8

Table 1: Candidate retrieval results at the UMLS CUI level on the official training/validation set (± label corrections) and the generalization set (R = Recall@k, %).

An evaluation of each retrieval component in isolation is presented in Appendix D.

5.2 Candidate Re-ranking

Table 2 reports the official test results for our system versus the shared-task reference baseline, which used edit-distance-based similarity. The large gap indicates that lexical similarity alone is insufficient for normalization from patient-written ADE mentions.

Model	A@1	A@5	A@10
Our system	39.8	78.3	85.5
Task reference baseline	12.1	13.3	16.9

Table 2: Official test set performance at the MedDRA LLT level (A = Accuracy@k, %).

The relatively wide difference between Accuracy@1 and Accuracy@5, with a smaller gain from Accuracy@5 to Accuracy@10, reflects a challenge of LLT-level discrimination. Many top 10 candidates are near-synonyms or surface-level variants. For example, in Figure 1, six of the top 10 candidates from our system are grouped under UMLS CUI C0080274 (urinary retention). It is possible that institution- or coder-specific preferences may be required to guide selection in these situations.

Ablations on the generalization set (Appendix F) at the re-ranking stage show that removing the LLM re-ranker substantially reduces performance, underscoring the value of leveraging its internal knowledge and providing guideline-based instructions and examples.

6 Conclusion and Future Work

We present a pipeline for normalizing patient-authored ADE mentions to MedDRA that tied for first place on the shared task, substantially outperforming a lexical baseline. In practice, the system reliably shortlists appropriate LLTs, but its Accuracy@1 of 39.8% remains insufficient for fully automatic normalization without human oversight.

Future improvements include fine-tuning a bi-encoder as a lower-latency alternative to the cross-encoder, exploring alternative LLMs, and adding dataset- or institution-specific tie-break conventions to the system prompt.

7 Limitations

Our work has several limitations. Although we include multiple datasets, the system's portability to other writing styles or terminologies is uncertain. While feasible on consumer hardware, throughput constraints and API costs from the multi-stage pipeline may hinder large-scale normalization.

References

Nestor Alvaro, Yusuke Miyao, and Nigel Collier. 2017. Twimed: Twitter and pubmed comparable corpus of drugs, diseases, symptoms, and their relations. *JMIR Public Health and Surveillance*, 3(2):e24.

Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. COMETA: A corpus for medical entity linking in the social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3122–3137, Online. Association for Computational Linguistics.

Maksim Belousov, William G. Dixon, and Goran Nenadic. 2019. MedNorm: A corpus and embeddings for cross-terminology medical concept normalisation. In *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 31–39, Florence, Italy. Association for Computational Linguistics.

Peter Beninger. 2018. Pharmacovigilance: An overview. *Clinical Therapeutics*, 40(12):1991–2004.

Peter Beninger. 2020. Signal management in pharmacovigilance: A review of activities and case studies. *Clinical Therapeutics*, 42(6):1110–1129.

Olivier Bodenreider. 2004. The unified medical language system (umls): Integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–D270.

Egon G. Brown, Louise Wood, and Sue Wood. 1999. The medical dictionary for regulatory activities (meddra). *Drug Safety*, 20(2):109–117.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

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*, page 104744.

Dina Demner-Fushman, Sonya E. Shooshan, Laritza Rodriguez, Alan R. Aronson, Francois Lang, Willie Rogers, Kirk Roberts, and Joseph Tonning. 2018. A dataset of 200 structured product labels annotated for adverse drug reactions. *Scientific Data*, 5(1):180001.

Luyu Gao, Zhuyun Dai, Tongfei Chen, Zhen Fan, Benjamin Van Durme, and Jamie Callan. 2021. Complement lexical retrieval model with semantic residual embeddings. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 – April 1, 2021, Proceedings, Part I*, page 146–160, Berlin, Heidelberg. Springer-Verlag.

Su Golder, Gill Norman, and Yoon K. Loke. 2015. Systematic review on the prevalence, frequency and comparative value of adverse events data in social media. *British Journal of Clinical Pharmacology*, 80(4):878–888.

Su Golder, Karen O'Connor, Yunwen Wang, Ari Klein, and Graciela Gonzalez Hernandez. 2024. The value of social media analysis for adverse events detection and pharmacovigilance: Scoping review. *JMIR Public Health and Surveillance*, 10:e59167.

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.

Jebran Khan, Kashif Ahmad, Senthil Kumar Jagatheesaperumal, and Kyung-Ah Sohn. 2025. Textual variations in social media text processing applications: challenges, solutions, and trends. *Artificial Intelligence Review*, 58(3):89.

Saar Kuzi, Mingyang Zhang, Cheng Li, Michael Bendersky, and Marc Najork. 2020. Leveraging semantic and lexical matching to improve the recall of document retrieval systems: A hybrid approach. *arXiv preprint arXiv:2010.01195*.

Nut Limsopatham and Nigel Collier. 2016. Normalising medical concepts in social media texts by learning semantic representation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1014–1023, Berlin, Germany. Association for Computational Linguistics.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. 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*, pages 4228–4238, Online. Association for Computational Linguistics.

MedDRA MSSO. 2025. What's New: MedDRA Version 28.1. Technical Report 001274, MSSO. PDF.

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.

Patricia Mozzicato. 2009. Meddra: An overview of the medical dictionary for regulatory activities. *Pharmaceutical Medicine*, 23(2):65–75.

Dimitra Pappa and Lampros K. Stergioulas. 2019. Harnessing social media data for pharmacovigilance: A review of current state of the art, challenges and future directions. *International Journal of Data Science and Analytics*, 8(2):113–135.

Abeed Sarker, Maksim Belousov, Jasper Friedrichs, Kai Hakala, Svetlana Kiritchenko, Farrokh Mehryary, Sifei Han, Tung Tran, Anthony Rios, Ramakanth Kavuluru, Berry de Bruijn, Filip Ginter, Debanjan Mahata, Saif M. Mohammad, Goran Nenadic, and Graciela Gonzalez-Hernandez. 2018. Data and systems for medication-related text classification and concept normalization from twitter: Insights from the social media mining for health (smm4h)-2017 shared task. *Journal of the American Medical Informatics Association*, 25(10):1274–1283.

U.S. National Library of Medicine. 2025. Umls knowledge sources [dataset on the internet]. Bethesda (MD): National Library of Medicine (US); Release 2025AA. Released 2024 May 6 [cited 2025 Oct 25].

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2024. Jasper and stella: Distillation of sota embedding models. arXiv preprint. ArXiv:2412.19048.

Richard C. Zink, Rebecca Lyzinski, and Geoffrey Mann. 2025. Aggregation of adverse event terms for signal detection and labeling in clinical trials. *Drug Safety*, 48(6):595–606.

Maryam Zolnoori, Kin Wah Fung, Timothy B. Patrick, Paul Fontelo, Hadi Kharrazi, Anthony Faiola, Nilay D. Shah, Y. Shirly Shirley Wu, Carolyn E. Eldredge, Jie Luo, Mike Conway, Jie Zhu, So-Young K. Park, Kun Xu, and Hamid Moayyed. 2019. The psytar dataset: From patients generated narratives to a corpus of adverse drug events and effectiveness of psychiatric medications. *Data in Brief*, 24:103838.

A Hardware Details

All experiments were conducted on Windows 10 with a single NVIDIA RTX 4070 Ti (12 GB VRAM) and 32 GB system RAM.

B TF-IDF Configuration

Text is lowercased, punctuation is removed, and start- and end-markers are added at word boundaries. We use a character-level `TfidfVectorizer` (2-5 character n-grams) with sublinear term frequency, smoothed inverse document frequency, L_1 normalization, and no vocabulary pruning.

C Cross-Encoder Configuration

For training, we used the LambdaLoss objective with the NDCG Loss2++ weighting scheme and $k = 20$. We trained for up to 20 epochs with early stopping on generalization set mean average precision (converged at epoch 3). Additional hyperparameters are listed in Table 6.

Component	Setting
Max sequence length	128
Train batch size	128
Optimizer	AdamW (fused)
LR schedule	Cosine
Learning rate	3×10^{-5}
Weight decay	0.01
Warmup ratio	0.1

Table 6: Hyperparameters used for cross-encoder fine-tuning.

D Retrieval Variant Results

We compared the performance of individual retrieval components with the final hybrid approach on the generalization set at the UMLS CUI level (Table 3). All dense retrievers outperformed TF-IDF alone, reflecting their ability to handle non-standard wording and phrasing. The hybrid approach yielded the best overall coverage (highest Recall@all), although it exhibited lower recall at smaller values of k. `stella_en_400M_v5` achieved the strongest single-model performance, outperforming the domain-adapted `SapBERT` model.

E Cross-Encoder Results

Table 4 reports CUI-level retrieval performance before and after cross-encoder re-ranking on the generalization set. The cross-encoder improved mid- and large-k recall by rescuing tail cases but slightly worsened performance at very small k. Despite performing similarly to the initial retrieval results, its ability to provide unified scores for candidate filtering increased its utility.

F Re-ranking Stage Ablations

Ablations in Table 5 isolate each component's contribution at the MedDRA LLT level. Filtering out non-current MedDRA terms reduces recall, although this practice aligns with MedDRA guidelines. Adding the cross-encoder yields a pattern of performance difference similar to that observed at the CUI level (Table 4). The higher Recall@300 without cross-encoder re-ranking is due to a lack of score threshold filtering.

G LLM Configuration

We queried Gemini 2.5 Pro via its API with dynamic thinking enabled, temperature $= 0.1$, top-$p = 0.9$, and top-$k = 40$.

Variant	Recall@10		Recall@50		Recall@100		Recall@all	
	Value	Δ	Value	Δ	Value	Δ	Value	Δ
Dense+Sparse Hybrid	64.4	0.0	82.7	0.0	87.4	0.0	**93.8**	0.0
Dense Ensemble Alone	67.9	3.5	78.6	−4.1	85.0	−2.4	93.4	−0.4
stella_en_400M_v5	**77.6**	13.2	**88.3**	5.6	**91.4**	4.0	91.4	−2.4
BGE_large_medical	73.5	9.1	87.7	5.0	90.1	2.7	90.1	−3.7
SapBERT	67.9	3.5	80.7	−2.0	84.2	−3.2	84.2	−9.6
Sparse Alone (TF-IDF)	53.9	−10.5	71.6	−11.1	77.2	−10.2	77.2	−16.6

Table 3: Retrieval performance on the *generalization set* at the UMLS CUI level, using Recall@k (%). Δ shows the absolute difference in Recall relative to the *Dense+Sparse Hybrid* (final system). "Dense Ensemble Alone" is the union of the three dense models (no TF-IDF). Single-model rows use only that retriever. Recall@all measures recall when taking all candidates returned by that variant. The highest-scoring variant for each metric is bolded.

Model	Recall@1	Recall@5	Recall@10	Recall@50	Recall@100
Original	37.2	56.4	64.4	82.7	87.4
Cross-encoder re-rank	35.0	58.8	67.0	83.4	89.7

Table 4: CUI-level retrieval before and after cross-encoder re-ranking (Recall@k, %).

Model	Recall@1	Recall@5	Recall@10	Recall@50	Recall@100	Recall@300
CE + Filtering + LLM Re-rank	**35.5**	**70.2**	**80.5**	—	—	—
CE + Filtering	23.4	50.0	60.6	83.7	87.6	92.2
CE	24.1	50.7	61.0	**84.0**	**89.7**	**94.7**
Filtering without CE	25.5	55.0	61.4	78.7	86.9	93.6

Table 5: Ablations at the MedDRA LLT level on the generalization-set subset of mentions with sentence context available (i.e. shared-task data only). Recall@k (%). Dashes (—) indicate not applicable. **CE** = cross-encoder re-rank. **Filtering** = exclusion of non-current, non-LLT/PT terms. The highest scoring condition for each metric is bolded.

System prompt The system prompt used for each mention was as follows:

> You are a clinical coding assistant specializing in MedDRA. Your job is listwise re-ranking:
> given (a) one short adverse event (AE) mention/verbatim ("MENTION_TEXT"), (b) the same mention within its original text, wrapped in <mention>...</mention> tags ("MENTION_IN_CONTEXT"), and (c) a set of candidate MedDRA code DESCRIPTIONS, return ONLY a JSON object with 1-based indices of the TOP 10 most relevant candidates in DESCENDING order of relevance.
>
> **## INPUT FIELDS**
> - MENTION_TEXT: the exact AE span only (the 'verbatim').
> - MENTION_IN_CONTEXT: the full user text surrounding the mention, with the AE span wrapped in <mention>...</mention>.
> - CANDIDATES: list of candidate MedDRA LLT/PT DESCRIPTIONS, 1-based.
>
> **## WHAT TO OPTIMIZE**
> Re-rank candidates to the best-fitting MedDRA Lowest Level Term (LLT) for MENTION_TEXT, following MedDRA Term Selection principles. Use MENTION_IN_CONTEXT only to DISAMBIGUATE the span (e.g., body site, finding vs. disorder, intended meaning), not to introduce additional reportable concepts.
>
> **## HARD RULES**
> - Use ONLY the provided candidates. Do NOT invent or rewrite terms.
> - While an attempt was made to rank candidates by relevance, the initial order may be arbitrary. Re-rank and consider ALL candidates, even ones in the middle and end of the list.
> - Focus on coding exactly what is reported in the span; do NOT add unmentioned diagnoses, causality, temporality, severity, or etiology.
> - Interpret lay language, misspellings and slang using best medical judgement. Text is sourced from patient forums.
> - Do not "up-normalize" from lay to medical: if both lay and medical LLTs express the same concept at the same specificity, prefer the LLT that best matches the reported wording in MENTION_TEXT.
> - Singular vs plural candidates: If singular/plural variants imply different medical concepts, choose the one that best represents the span's meaning rather than the grammar.
> - Spelling variants (UK/US) and word order variants: if both are viable and equally specific, prefer the variant/word order that most closely matches MENTION_TEXT.

- Prefer a single LLT that matches a combined concept when a suitable combination term is present (e.g., “Itchy rash”). Otherwise, represent the concepts individually; “split” only when no single candidate captures the distinct concepts.
- Body site vs medical event: if an “event + site” LLT exists and context shows the site is integral or prevents confusion, prefer that combined term. Otherwise prioritize the EVENT over the site.
- Definitive diagnosis with signs/symptoms in the same span: prioritize the diagnosis (do not double-code signs/symptoms contained within the diagnosis unless the diagnosis is uncertain).
- Investigations:
- If result direction is unambiguous (e.g., numeric with units below/above range), prefer a directional result LLT (e.g., “Blood glucose low” / “Potassium increased”).
- If ambiguous or the text and result direction conflict, prefer a non-directional abnormal result (e.g., “Glucose abnormal”).
- Prefer investigation-result terms (e.g., “Low blood glucose” over disease diagnoses (e.g., “Hypoglycemia”) when only a test result is reported.
- If the span clearly implies multiple distinct reportable concepts and no single candidate captures them, rank strong candidates for each concept near the top (“split” behavior).
- Pre-existing conditions: if the span reflects an event on a background of an unchanged pre-existing condition, code the event (unless a single appropriate combination term exists). If the span reflects an event that alters (e.g., aggravates) a pre-existing condition, prefer a corresponding LLT.
- Neoplasms: do not infer malignancy unless explicitly stated.
- Suicide/self-harm/overdose: do not assume overdose means suicide, or that ideation implies action; code exactly what is stated (e.g., Accidental overdose vs Intentional overdose).
- When MENTION_IN_CONTEXT conflicts with the MENTION_TEXT: the span wins. Use surrounding context only to clarify the span’s meaning.
- Consider the nuances between Impairment / abnormality / disease / disorder:
- Use investigation-result terms for test findings (increased/decreased/abnormal); do not infer a disorder from a result.
- Use disorder/disease terms only when a clinical condition is actually reported.
- Use impairment/disability/person-status terms when that is what is reported, not a medical condition.
- Adjective Use: Prefer the adjective form, e.g., “cardiac” or “hepatic” instead of the noun (e.g., “heart” or “liver”). The exceptions are when the term is not normally stated as such in common practice (e.g., “heart attack” over “cardiac attack”).

TIE-BREAKERS (apply in order)

1) Exact/near-exact semantic match to MENTION_TEXT that best captures the reporter’s words or intended meaning (including number, spelling variant, and word order).
2) Appropriateness of the description’s category (e.g., sign/symptom vs disorder vs personal circumstance) to what the span reports.
3) Appropriate combination or event+site LLT when justified by the span/context without adding unmentioned information.
4) Higher clinical specificity over generic wording, without assuming site/etiology not in the span/context.
5) If two candidates are the same concept but one is “NOS”, prefer the non-NOS candidate.
6) If two candidates are equivalent and differ only in word order or spelling variant, prefer the one closest to MENTION_TEXT; if still tied, prefer the UK spelling option and/or the earlier 1-based index mention.
7) If still tied after all above, prefer the candidate with the earlier 1-based index.

OUTPUT

- Return ONLY JSON with schema: { "ranking": [i1, i2, . . . , iK] }
- Indices are 1-based and unique. K = min(10, number of candidates).
- Do NOT output any text outside the JSON.

QUALITY CHECKS (after ranking)

- Indices must be within range.
- No unreported information is present in the top candidate.