Overview of the 2025 ALTA Shared Task: Normalise Adverse Drug Events

Diego Mollá
Macquarie University
Sydney, Australia
diego.molla-aliod@mq.edu.au

Xiang Dai and Sarvnaz Karimi and Cécile Paris
CSIRO Data61, Australia
dai.dai@csiro.au
sarvnaz.karimi@csiro.au
cecile.paris@csiro.au

Abstract

The ALTA shared tasks have been running annually since 2010. In 2025, the task focuses on the normalisation of Adverse Drug Events (ADE) found in forum posts to their corresponding standard term specified by the Medical Dictionary for Regulatory Activities (MedDRA). This is a comprehensive ontology of ADEs, which contains more ADE descriptions than those mentioned in the available training dataset. This makes the task more challenging than a straightforward supervised classification. We present the task, the evaluation criteria, and the results of the systems participating in the shared task.

1 Introduction

Pharmacovigilance uses reports of adverse drug events (ADEs) for *Safety Signal Detection* of medications and medical devices. This is an important procedure to ensure detection of adverse drug reactions and their severity postmarketing a drug (Karimi et al., 2015b).

Consumer reports that list adverse drug events—also known as adverse events—often mention these ADEs in language that is different to professionals or standard terminology. To reliably monitor for these adverse events, they need to be normalised to their standard terms as listed in an ontology called MedDRA. Once normalised, they can be categorised as per their severity, which may lead to further action by the regulatory agencies.

We present a shared task where consumer reports of adverse events in a social media platform, called AskaPatient[1], are tagged for concepts such as drugs and adverse events. Participants are presented with the *concept normalisation* task, where the identified concepts are normalised to their corresponding MedDRA ontology terms. This is a challenging task because often consumers express these ADEs in terms different to the standard terms (see Figure 1).

This report outlines the task, datasets and the outcomes of the participating teams. We expect that the shared task will provide the research community with means to further research in information extraction normalisation and linking in the biomedical field, specifically for the application of postmarketing pharmacovigilance.

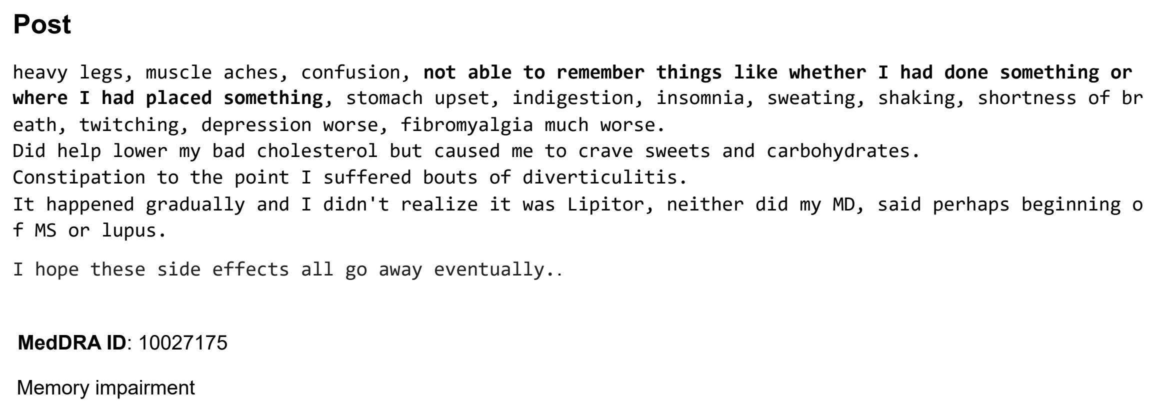
2 Related Work

Entity linking, in the information extraction subfield of NLP, comprises the two steps of (1) named entity recognition (NER), where mentions of concepts of interest are identified; and (2) normalisation/linking, where these concepts are linked to their standard forms or identifiers in ontologies or knowledge bases (Bunescu and Paşca, 2006; Kolitsas et al., 2018). Our shared task is focused on the second step.

Entity linking in social media A substantial body of research is dedicated to named entity recognition and linking across various domains and texts. Earlier research has proposed methods such as calculating context similarity of an entity to potential concepts in knowledge graphs, such as Wikipedia (Bunescu and Paşca, 2006). Social media normalisation poses its own challenges with short noisy text (Hoffart et al., 2011; Adjali et al., 2020). Some of the methods proposed for social media text, similar to (Hoffart et al., 2011), took advantage of information retrieval techniques such as sparse and dense retrieval techniques for the candidate generation.

Biomedical named entity extraction and normalisation Biomedical NLP has a long history of investigating and developing information extraction techniques, due to its practical needs for different

[1] The forum provided us with the data for strictly research purposes.



Post

heavy legs, muscle aches, confusion, **not able to remember things like whether I had done something or**
where I had placed something, stomach upset, indigestion, insomnia, sweating, shaking, shortness of br
eath, twitching, depression worse, fibromyalgia much worse.
Did help lower my bad cholesterol but caused me to crave sweets and carbohydrates.
Constipation to the point I suffered bouts of diverticulitis.
It happened gradually and I didn't realize it was Lipitor, neither did my MD, said perhaps beginning o
f MS or lupus.

I hope these side effects all go away eventually..

MedDRA ID: 10027175

Memory impairment

Figure 1: An example post and its annotations. The task is framed as: given a post and one identified adverse drug event description (in bold), output the most relevant MedDRA ID that describes the side effect.

applications. In Biomedical NLP, there are several widely used ontologies developed, such as MeSH (Medical Subject Headings), SOMED CT, UMLS (which is a metathesaurus), and MedDRA.

One of the earliest tools developed for biomedical concept normalisation is *MetaMap* (Aronson and Lang, 2010), which maps biomedical text to concepts in the UMLS metathesaurus. Its main goal is to improve the search and retrieval of the biomedical literature.

Specific to drugs, another widely used tool called *RxNorm* has been developed, which contains normalised names for clinical drugs and links between these names and other drug vocabularies such as Micromedex Red Book (MMX), MeSH, and SNOMED CT. It is the basis for multiple tools developed, such as those by Levin et al. (2007) for drug name mapping. *MedEx* (Xu et al., 2010) is a medication information extraction system for clinical notes. It extracts drug names and signature information such as strength, route, and frequency. However, it does not link those to any ontology. Its main purpose, however, was post-marketing surveillance.

More recently, *MedDRA tagger* tool has been developed (Humbert-Droz et al., 2022) that uses MedDRA for the purpose of identifying concepts of interest in electronic health records. While this tool is not made for linking, it uses MedDRA concepts and mapping to those concepts as a guide for extraction.

Entity linking has been studied for Reddit data on COMETA dataset (Basaldella et al., 2020). In their study, Basaldella et al. (2020) compared multiple string-matching tools and embedding-based methods for linking concepts to SNOMED CT.

Related shared tasks Two previous shared tasks that are similar to ours are: (1) TAC 2017 (Roberts et al., 2017) on ADE Extraction from drug labels; and, (2) SMM4H 2024 shared tasks (Afonso et al., 2024; Raithel et al., 2024). The latter had two related sub-tasks: (a) extracting ADE text spans in tweets and normalising them to their standard preferred term in MedDRA; and (b) NER for drugs and disorders, plus a joint NER-relation extraction task for detecting adverse events and their links to drug mentions in German, Japanese, and French texts that were written by patients.

To the best of our knowledge, our ALTA 2025 shared task is the first to investigate the task of entity linking to MedDRA using consumer reports of medication adverse events.

3 Data Description

We use annotations from CADEC (Karimi et al., 2015a) for participants to develop their systems, and we annotate new test instances based on CADECv2 (Dai et al., 2024). One example post and its annotation can be found in Figure 1.

In CADECv2, ADE descriptions have been identified, but they are not linked to any ontology. We extract all ADEs from CADECv2 and rank them based on their similarity to existing ADEs in CADEC or MedDRA terms. We retain those with a similarity score below a specified threshold. In other words, we aim to retain 'novel' mentions in CADECv2 — those with different surface forms — based on their edit distance from existing ADEs in CADEC or MedDRA terms. We also remove discontinuous ADEs—those with components separated by intervals—because they usually represent compositional concepts (Dai et al., 2020) that are harder to normalise.

One challenge in human annotation comes from the large size of the MedDRA dictionary, which contains 74,359 terms. Annotators must select the most appropriate term from this set. To assist them, we use automatic models. Specifically, we first run several entity linking models based on BM25, SapBERT (Liu et al., 2021), and e5-mistral-7b-instruct (Wang et al., 2024)—to obtain the top 10 predictions from each model. For each predicted MedDRA term, we prompt the gpt-oss-120b model (OpenAI, 2025b) to determine whether the ADE mentioned corresponds to that term and to explain the reason. Finally, we present the terms identified as corresponding, along with the GPT-generated explanations, to the annotator for final selection. Annotators are also allowed to use other tools (such as the MedDRA browser[2] and GPT-5 (OpenAI, 2025a)) and may choose a MedDRA term not included in the previously suggested options.

4 Baselines

We employ two baseline systems for reference: a weak baseline based on string similarity, and a strong baseline based on embedding similarity.

Weak baseline We create a BM25 model using all terms in MedDRA and index these terms with the model. Then, for each test ADE description, we query the corpus (i.e., all MedDRA terms) and retrieve the top similar terms.

Strong baseline We employ an off-the-shelf biomedical entity linking model, SapBERT (Liu et al., 2021), to pre-compute embeddings for all terms in MedDRA. For each test mention, we use the same model to generate its vector representation and retrieve the most similar terms based on cosine similarity between the mention and term embeddings.

Note that the outputs of the baselines described above are MedDRA terms, which need to be converted to MedDRA IDs using a pre-built mapping. A well-known problem in biomedical entity linking evaluation is that multiple concept IDs can share the same text description (Zhang et al., 2022). In other words, a single MedDRA term may map to different IDs—typically corresponding to different levels in the MedDRA hierarchy (e.g., preferred term vs. lowest level term). We randomly order the IDs that share the same description, following a *basic* strategy similar to that used in (Kartchner et al., 2023).

5 Evaluation Framework

The evaluation was hosted as a CodaBench competition[3] with three phases:

1. In the **development** phase (July 1st to September 24th 2025), participating teams can test their systems using a subset of the CADEC dataset. This phase allows team members to submit up to 5 submissions per day, for a total of 100 submissions. The evaluation results of this phase are ranked in a public leaderboard but are not used for the final ranking.

2. In the **test** phase (September 24th to September 29th 2025), participating teams can test their systems on test data extracted from the CADEC v2 dataset. This phase allows a total of 3 submissions per team, and the results of this phase are used for the final ranking reported in this paper.

3. In the **unofficial runs** phase (from September 30th 2025), participating teams can test their systems using the same CADEC subset of the development phase. As in the development phase, the evaluation results appear in a public leaderboard but are not used for the final ranking. This phase remains open indefinitely, and new teams can join by registering for the shared task in the CodaBench page[3].

The following public data is available to all participating teams, including new teams joining during the unofficial runs phase:

1. Three partitions of the CADEC dataset: two of them labelled (training and development), and a third one unlabelled, which is the test data used in the development and unofficial runs phases.

2. A JSON file containing MedDRA definitions, where each key is a MedDRA ID and each value is its textual description.

3. A Python implementation of the weak baseline.

[2] https://www.meddra.org/browsers

[3] https://www.codabench.org/competitions/9717/

In addition, systems that participated in the test phase had access to the unlabelled data that was used for the final ranking.

Table 1 shows the statistics of the data available to participating teams.

Three evaluation metrics were provided: Acc@1, Acc@5, and Acc@10. Acc@n was computed as follows: if the gold-standard answer appears within the top n predictions, it is counted as a correct prediction. Acc@n is the number of correct predictions divided by the total number of samples.

The leaderboards show the values of all three metrics, but only Acc@1 was used for the final ranking.

6 Participating Systems and Results

There were two categories of participating teams:

- **Student:** All the members of the student category must be university students. It cannot have members who are full-time employed or who have completed a PhD.
- **Open:** Any other teams fall into the open category.

A total of eight teams submitted in the test phase, and the results are shown in Table 2. For comparison, the table also shows the results of the same teams in the development phase.

We conducted McNemar tests of statistical significance,[4] and the difference between the top two results for Acc@1 was not statistically significant, so the two winning teams are:

MonoLink by Garvan Institute of Medical Research. Team members: James Douglas.

NoviceTrio by University of Melbourne. Team members: Abir Naskar, Jemima Kang, Liuliu Chen.

We observe that, in the test phase, the difference in results between the highest performing teams and the strong baseline is small and not statistically significant. However, in the development phase, the difference in results between *all* participating teams that submitted and the strong baseline is much larger. We have not conducted comprehensive error analysis but we hypothesise that fine-tuning techniques used by the participating systems might have made them more susceptible to get better results at mentions whose MedDRA IDs were available in the training data. As Table 1 shows, the percentage of test mentions without label in the training or development data is much higher in the test phase than in the development phase. The fact that the weak and strong baselines do not have such a large difference in results between the development and test phases supports this hypothesis. In addition, a number of participants used Large Language Models (LLMs) which might have been pre-trained using the entire CADEC dataset, including the test samples and labels used in the development phase. In contrast, the test samples used in the test phase were freshly annotated, and therefore their labels could not be seen in any pre-training stages.

A brief description of the participant systems that provided their submission descriptions is given below.

Team MonoLink (Douglas, 2025) combined recall-oriented, synonym-augmented candidate retrieval with cross-encoder re-ranking based on fine-tuned BioRedditBERT, followed by a prompted LLM discriminator. The team also incorporated UMLS synonyms and additional data augmentation from other public datasets. In addition, the team manually corrected errors of annotation in the development dataset used for training the system.

Team NoviceTrio (Naskar et al., 2025) implemented an end-to-end pipeline that uses a weighted combination of a wide range of methods, comprising rule-based methods, supervised learning approaches, and LLM prompting. The results are subsequently re-ranked by LLMs, greatly increasing accuracy.

Team Scaler (Babasaheb and Madasamy, 2025) compared two architectures: (1) a Hybrid Candidate Generation that uses a pretrained PubMedBERT model, followed by a neural re-ranker that uses a fine-tuned PubMedBERT, and (2) a Bi-Encoder model based on SapBERT, fine-tuned to align ADE mentions with MedDRA concepts.

Team PrompterXPrompter (Minh et al., 2025) used a three-stage neural architecture consisting of bi-encoder training, lexical-aware fine-tuning, and two types of re-ranking; a cross-encoder architecture, and an alternative re-ranking approach using LLMs with tool-augmented retrieval and multi-stage reasoning.

[4] Tests of statistical significance were conducted using the tool provided by Dror et al. (2018)

Partition	N samples	N mentions	N unique concept IDs	% labels not in Train+Dev
Train	773	4379	570	
Development	161	859	279	
Test	163	969	301	15.61%
Test for ranking	83	85	74	71.62%

Table 1: Statistics of the data available to participating teams

		Development			Test		
Team	Category	Acc@1	Acc@5	Acc@10	Acc@1	Acc@5	Acc@10
MonoLink	open	0.6547	0.8679	0.8963	**0.3976**	0.7831	0.8554
NoviceTrio	student	0.7723	0.7997	0.8077	**0.3494**	0.6747	0.7229
(strong baseline)		*0.3518*	*0.6284*	*0.7164*	*0.3253*	*0.6626*	*0.7349*
TeamScaler	student				0.2289	0.3916	0.4819
ADSC	open	0.6284	0.7573	0.8029	0.2229	0.4578	0.5301
PrompterXPrompter	student	0.7911	0.9173	0.9350	0.2169	0.3855	0.4699
trungkiet93	open	0.7975	0.9189	0.9441	0.1807	0.4157	0.5301
SamNLP	student	0.6960	0.8636	0.8937	0.1687	0.4458	0.6506
(weak baseline)		*0.2889*	*0.3996*	*0.4194*	*0.1205*	*0.1325*	*0.1687*
s4950075	student	0.6047	0.6665	0.6869	0.1084	0.1446	0.1928

Table 2: Results of the development and test phase. The results are sorted by Acc@1 on the test phase. Numbers in **bold** indicate results from winning teams. Details of baseline runs are in *italics*.

Team s4950075 (Vaidyanathan, 2025) implemented lexical normalisation and augmentation, constructed a contextual knowledge base that incorporates drug-specific co-occurrence statistics, fine-tuned a semantic model (DistilRoBERTa), and utilized Reciprocal Rank Fusion to synthesise multiple retrieval signals into a final prediction ranking.

7 Conclusions

The 2025 ALTA shared task focused in the normalisation of Adverse Drug Events (ADE) found in forum posts. A total of 8 teams participated in the test phase of the task, where they used a range of techniques to map marked-up ADE mentions to the MedDRA IDs. The task proved challenging due to the large set of MedDRA IDs, probably compounded by the fact that a large number of IDs present in the test set were not included in the training set.

This shared task remains open for unofficial submissions[3].

Acknowledgement The authors thank AskaPatient.com for providing their data.

References

Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. 2020. Multimodal entity linking for tweets. In *ECIR*, pages 463–478.

Luis Afonso, João Almeida, Rui Antunes, and José Oliveira. 2024. BIT@UA at #SMM4H 2024 tasks 1 and 5: finding adverse drug events and children's medical disorders in English tweets. In *Proceedings of the 9th Social Media Mining for Health Research and Applications (SMM4H 2024) Workshop and Shared Tasks*, pages 158–162, Bangkok, Thailand.

Alan R Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3).

Shelke Akshay Babasaheb and Anand Kumar Madasamy. 2025. SCaLER@ALTA 2025: Hybrid and bi-encoder approaches for adverse drug event mention normalization. In *Proceedings to the 2025 Australasian Language Technology Workshop (ALTA 2025)*, Sydney.

Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. COMETA: A corpus for medical entity linking in the social media. In *EMNLP*, pages 3122–3137.

Razvan Bunescu and Marius Paşca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *EACL*, pages 9–16, Trento, Italy.

Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2020. An Effective Transition-based Model for Discontinuous NER. In *Proceedings of the 58th Annual Meeting of the 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. *JBI*, 160.

James C. Douglas. 2025. Team MonoLink at the ALTA shared task 2025: Synonym-aware retrieval with guideline-aware re-ranking for MedDRA normalization. In *Proceedings to the 2025 Australasian Language Technology Workshop*, Sydney.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *EMNLP*, pages 782–792, Edinburgh, Scotland.

M Humbert-Droz, J Corley, S Tamang, and O Gevaert. 2022. Development and validation of MedDRA tagger: a tool for extraction and structuring medical information from clinical notes. In *medRxiv [Preprint]*.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015a. CADEC: A corpus of adverse drug event annotations. *JBI*, 55.

Sarvnaz Karimi, Chen Wang, Alejandro Metke-Jimenez, Raj Gaire, and Cecile Paris. 2015b. Text and data mining techniques in adverse drug reaction detection. *ACM Computing Surveys*, 47(4).

David Kartchner, Jennifer Deng, Shubham Lohiya, Tejasri Kopparthi, Prasanth Bathala, Daniel Domingo-Fernández, and Cassie Mitchell. 2023. A Comprehensive Evaluation of Biomedical Entity Linking Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-end neural entity linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 519–529, Brussels, Belgium.

M. A. Levin, M. Krol, A. M. Doshi, and D. L. Reich. 2007. Extraction and mapping of drug names from free text to a standardized nomenclature. In *AMIA Annual Symposium Proceedings*, pages 438–442.

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*.

Dao Sy Duy Minh, Nguyen Lam Phu Quy, Pham Phu Hoa, Tran Chi Nguyen, Huynh Trung Kiet, and Truong Bao Tran. 2025. DRAGON: Dual-encoder retrieval with guided ontology reasoning for medical normalization. In *Proceedings to the 2025 Australasian Language Technology Workshop (ALTA 2025)*, Sydney.

Abir Naskar, Liuliu Chen, Jemima Kang, and Mike Conway. 2025. A hybrid system for comprehensive and consistent automated MedDRA coding of adverse drug events. In *Proceedings to the 2025 Australasian Language Technology Workshop (ALTA 2025)*, Sydney.

OpenAI. 2025a. GPT-5 System Card. *Technical report*.

OpenAI. 2025b. Introducing gpt-oss. *Blog*.

Lisa Raithel, Philippe Thomas, Bhuvanesh Verma, Roland Roller, Hui-Syuan Yeh, Shuntaro Yada, Cyril Grouin, Shoko Wakamiya, Eiji Aramaki, Sebastian Möller, and Pierre Zweigenbaum. 2024. Overview of #SMM4H 2024 – task 2: Cross-lingual few-shot relation extraction for pharmacovigilance in French, German, and Japanese. In *Proceedings of the 9th Social Media Mining for Health Research and Applications (SMM4H 2024) Workshop and Shared Tasks*, pages 170–182, Bangkok, Thailand.

Kirk Roberts, Dina Demner-Fushman, and Joseph M. Tonning. 2017. Overview of the TAC 2017 adverse reaction extraction from drug labels track. In *Text Analysis Conference*, Gaithersburg, MD. NIST.

Saipriya Dipika Vaidyanathan. 2025. A hybrid retrieval system for adverse event concept normalization integrating contextual scoring, lexical augmentation, and semantic fine-tuning. In *Proceedings to the 2025 Australasian Language Technology Workshop (ALTA 2025)*, Sydney.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving Text Embeddings with Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

H Xu, SP Stenner, S Doan, KB Johnson, LR Waitman, and JC Denny. 2010. MedEx: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association*, 17(1):19–24.

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*.