A Framework for LLM-Guided and ISO 5338-Aligned Data Engineering in Medical Device Adverse Event Reporting *

Muhammad Muaz Zaffar Haider Janjua
Roisin Loughran Niamh St John Lynch

April 26, 2026

Abstract

Medical device adverse event reporting systems, such as Manufacturer and User Facility Device Experience (MAUDE) database from the Food and Drug Administration (FDA), play a critical role in post-market surveillance and patient safety monitoring. However, the accurate assignment of standardized Adverse Event Terminology (AET) codes, particularly the globally harmonized International Medical Device Regulators Forum (IMDRF) codes, remains a labour-intensive task prone to variability in human judgment. This study presents a Large Language Model (LLM)-guided LLM-agreement filtering methodology for creating a high-confidence benchmark dataset from MAUDE records. By requiring agreement between human-annotated IMDRF codes and predictions from state-of-the-art LLMs at Level 2 hierarchical granularity, a low-ambiguity subset of 1,728 device problem records and 1,224 health effect records from an initial pool of 10,000 adverse events is established. The approach aligns the entire data pre-processing pipeline with ISO/IEC 5338:2023 AI data engineering principles, implementing modular Agent Skills that lays foundation for reproducibility, traceability, and quality assurance. Moreover, the open-source GPT-OSS-120B reasoning model is evaluated on this curated dataset, revealing significant challenges in cross-model alignment: the model achieves only 59.49% consensus at Level 1 for device problems and drops to 23.04% at Level 3, highlighting the difficulty of reconciling diverse textual inputs with rigid regulatory terminologies in long-context scenarios. These findings underscore the need for specialized NLP workflows beyond simple zero-shot prompting and establish a rigorous benchmark for future automated adverse event coding research.

*This publication has emanated from research jointly funded through Lero - the Research Ireland Centre for Software under Grant # 13/RC/2094_2, the Horizon Europe Framework Programme under project Recon4IMD (Grant # 101080997), and Taighde Éireann – Research Ireland under Grant # 21/FFP- A/9255. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

1 Introduction

Raw data is noise; coded data is actionable. The Manufacturer and User Facility Device Experience (MAUDE) database [19] is an important resource for the post-market surveillance of medical devices and is relied upon by medical device manufacturers, Food and Drug Administration (FDA) analysts, and clinical researchers to monitor patient safety. This database hosts the description of post-market adverse events that occur in the use of medical devices and assigns Adverse Event Terminology (AET) codes to standardize the description of the adverse event. This standardization through the AET coding is important for several reasons:

- It creates cross-border harmony and safety. For example, a "battery failure" in an insulin pump ought to be interpreted in the same way whether it is used in the United States of America (USA) or the European Union (EU).
- It alarms the FDA to cases where the same AET codes are reported multiple times in a defined period.
- If a unique fault has occurred in the use of a medical device but it is only described in an unstructured way without a standardized identifiable code, it might not be highlighted in time for safety recalls and trend analysis.
- It is legally mandated under FDA 21 CFR Part 803 to submit complete and timely reports [3]. Therefore, critical events must be apparent in the unstructured adverse event descriptions of MAUDE database.

The adverse events occur with the medical devices and the manufacturers of these medical devices are required to code these adverse events for the database. These manufacturers require an expert from their regulatory affairs or quality assurance department, if they have one. To automate this coding process, an ontology generator (such as proposed in [17]) can be utilized. This however, does not assist a non-expert in its reasoning to choose the right AET code. At a deeper level, Natural Language Processing (NLP) workflows can be used to assist coders as are done widely for the International Code of Diseases (ICD) coding [2, 10, 15]. This will not just lift the burden of correct coding based on the unstructured reports, but can also help guide the right approach to report these events.

Similar to ICD coding, researchers have also used NLP and machine learning methods for the automated coding of MAUDE adverse events to enable the downstream analysis tasks. For example, a benchmark corpus of 935 MAUDE reports is developed earlier annotated with key entities to facilitate supervised information extraction [22]. However, their annotated dataset is designed for supervised machine learning comparison. ClinicalBERT was fine-tuned to automatically identify health information technology-related adverse events, using a dataset of over 3,000 MAUDE adverse event reports labelled by domain experts [12]. However, their NLP technique requires fine-tuning of their model that

means no guarantee of how the model will work on generally any MAUDE description. Similarly, machine learning models (including CNNs) were trained to detect IT-related device events in MAUDE, achieving approximately 87%–90% accuracy ($F_1 \approx 0.87$) in classifying reports of that subtype [20]. Though such studies prove the need of NLP-workflows, the current rapid LLM improvements justify to study the LLM performance in zero-shot settings so that the system is reliable on coding the adverse events written by a diverse community of manufacturers.

As the current open-source LLMs utilise long context windows and display strong reasoning capabilities, they could be employed to assist coders in coding these adverse events. Benchmarking MAUDE database in its current form for evaluating the NLP workflow requires the ground-truth to be trustworthy. However, there are variations in the MAUDE AET coding as studied before (e.g., see [9,14]), which shows that it can be prone to the errors in human judgements. This also leads to the situation where there could be more than one code that can fit an adverse event description depending on human perspectives coming from diverse clinical, demographic, and functional situations.

This study proposes to benchmark a trustworthy subset of the MAUDE data so that the NLP and agentic workflows can be readily evaluated on it. To prepare this dataset, an extract-transform-load (ETL) workflow combined with LLM was proposed earlier to streamline how adverse events are identified and coded [16]. While the study utilized a standard ETL pipeline, this study's methodology aligns the data life cycle with ISO/IEC 5338:2023 [7], which is not specific to healthcare standards only. We propose to transition the process from experimental prompting to a standardized life cycle process, treating AI model engineering as a formal extension of traditional software implementation. The MAUDE pre-processing workflow is defined as a versioned set of modular Agent Skills [21]. Each skill is a self-contained package of instructions and rule-based logic that an AI system loads on-demand to execute a specific stage of the data pipeline.

The current FDA's AET codes are mapped to the internationally standardized AET coding developed by IMDRF, referred to as the IMDRF AET codes in the rest of the document. This mapping is done to harmonize the coding system internationally across the USA and Europe. In Europe, the use of IMDRF coding is mandated by the mandatory reporting form i.e., Manufacturer Incident Report (MIR) form [13].

To cross-check and filter an unambiguous set of the adverse event descriptions and their coding, a state-of-the-art LLM model trained on world knowledge (400B+ parameters) is proposed to be used along with the actual MAUDE codes, and then only those codes and their descriptions are to be selected where the LLM and MAUDE-given manual codes are the same. This leverages the MAUDE's data abundance and ensures that the output set contains a more confident, unambiguous set of adverse event descriptions. At the end of this study, the OpenAI's open-source reasoning, GPT-OSS 120B [1], is tested to close the study's loop and open venues for further improvements in the efficiency, interpretability, and wider adoption of the NLP-workflows development.

The rest of the paper is organized as follows: Section 2 first introduces ISO 5338 data engineering process and IMDRF coding. Section 3 explains the data pre-processing steps and their alignment with ISO 5338 data engineering principles. Section 4 applies the GPT-OSS-120B model on the dataset and reports the results. Section 5 concludes the paper.

2 Background Theory

2.1 AI Data Engineering Process in ISO/IEC 5338

ISO/IEC 5338:2023 [6] defines an AI data engineering process that encompasses all activities needed to acquire, curate, and prepare data for machine learning and outlines the requirements to be in a rigorous and traceable manner. Key tasks include: (a) acquiring or selecting raw data (e.g., gathering relevant records from databases or sensors); (b) conducting data labelling if needed to assign ground-truth outputs to inputs; (c) analysing and exploring the data to understand its content, domain patterns, and potential issues; (d) analysing data quality – assessing aspects like completeness, bias, and noise to guide cleaning and correction; (e) documenting data lineage and provenance – recording the data's origins, processing history, and context to support error tracing and maintenance; and (f) cleaning, merging, and transforming the data into the desired format. These steps include operations such as deduplication, filtering out irrelevant or erroneous entries, standardizing formats, and encoding data according to the AI model's requirements. We followed requirements from this standard in the paper to process data.

2.2 IMDRF Adverse Event Codes

IMDRF codes act as the international unified language of adverse events associated with the use of medical devices [4]. It describes an adverse event in a sentence and associate a code with it (like E0602). This coding is hierarchical based on increasing level of specificity from Level 1 (highest) to Level 3 (lowest). The coding is divided into seven categories (Appendices A to G), where Appendix A codes for medical device malfunction, Appendices E and F codes the harm/health effect on the patient/user. This study focuses on these three appendices, which in total are 1522 code. An example hierarchical code of Appendix A is shown in Fig. 1.

3 Data Preparation

3.1 MAUDE Data Pre-processing

Data Cleaning

To evaluate the ability of an NLP-workflow in classifying adverse event narratives into specific IMDRF codes, the initial dataset is comprised of the event de-

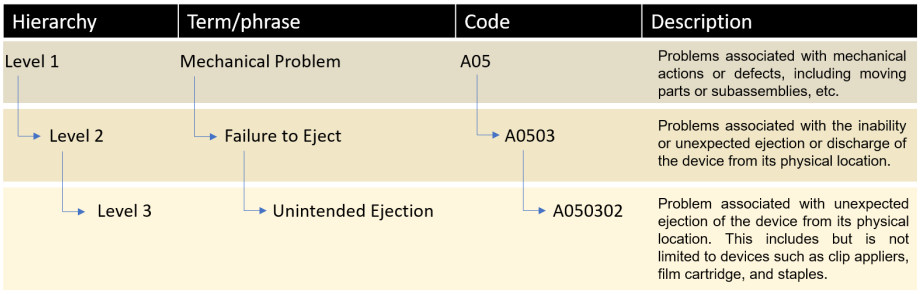

Figure 1: IMDRF Code Hierarchical Structure Example of Appendix A.

scription (that mostly contains the manufacturer's additional narrative) and the linked device problem and health effect codes. 10,000 records of adverse events are randomly extracted from the past 10 years ($2016 - 2025$) that contain device malfunction, death, and serious injury cases. These can be downloaded using the openFDA application programming interface [8] or directly by downloading the csv files based on a manual search criterion on the MAUDE website [19]. The reports are de-identified and already comply with Health Insurance Portability and Accountability Act (HIPAA) [18] and FDA guidelines. This curated dataset serves as the foundation for the pre-processing procedures described in the following text. After ensuring that no personally identifiable information exist, the raw MAUDE data is pre-processed to

1. remove redundant entries,
2. remove non-alphanumeric and non-punctuation-mark characters from event descriptions,
3. remove records without device problem or health effect codes, and
4. remove event descriptions coded as 'Insufficient Information', 'Adverse Event Without Identified Device or Use Problem', or 'Appropriate Term/Code Not Available' despite being given the AET code.

Conversion to IMDRF Codes

The current raw MAUDE database contains the FDA's Medical Device Report (MDR) adverse event codes instead of the IMDRF AET codes. The FDA has provided one-to-one mapping of their FDA codes with the IMDRF codes, that is utilized to re-label the dataset [5]. The pre-processing steps reduces the data records from 10000 to 2899.

LLM-Agreement

Due to the variation in human-written MAUDE AET coding among different device operators (as noted in [9, 14]), it is pertinent to cross-check whether

these adverse event descriptions point to the correct codes or not. This can be manually conducted by human experts going through thousands of records of clinically diverse situations to verify the correctness of IMDRF codes, which is logistically expensive to perform. On the other hand, instead of verifying the codes, an independent judgement is taken from a state-of-the-art LLM through its API, where the LLM is only prompted with the adverse event descriptions (without showing any AET code) to find the most suitable IMDRF code/s and then only those data records are moved forward on which both the LLM and MAUDE database have the same IMDRF codes. Instead of completely relying only on the MAUDE mappings or the state-of-the-art LLM, this LLM-agreement filter leaves a low-ambiguity subset of the MAUDE adverse event descriptions for the NLP-workflow automation and evaluation purposes. Also, getting the same code from two different perspectives means that the descriptions more clearly point to the code ensuring low ambiguity.

It is important to contextualise what the filter retains and what it discards, given the specific nature of MAUDE reporting. MAUDE adverse event reports are predominantly filed by manufacturers' regulatory affairs professionals, who are trained to use cautious, non-attributing language under FDA 21 CFR Part 803 requirements [3]. This produces a systematic pattern of linguistic minimisation in manufacturer-filed Medical Device Reports, as documented in the literature [9, 14]: narratives are deliberately hedged and non-causal (e.g., "the device may have contributed to") to protect the manufacturer from undue liability. As a consequence, many MAUDE narratives are not randomly ambiguous — they are *strategically* non-committal, making it genuinely difficult for any coder to assign a single IMDRF code with confidence regardless of their expertise. The LLM-agreement filter therefore does not merely select "easy" cases at random; rather, it identifies reports where the narrative, despite strategic regulatory language, is *reliably codeable* — i.e., where an independent LLM converges on the same code as the human coder. This is precisely analogous to the standard NLP dataset curation practice of filtering for inter-annotator agreement, which is widely accepted as a principled quality criterion rather than a source of bias. The retained benchmark thus represents the *reliably codeable* stratum of MAUDE, which is the appropriate target population for evaluating an NLP coding workflow. The filter also relies on a single model (Gemini Flash 3); future work should replicate this step with multiple LLMs and report inter-model agreement scores to further validate the retained subset.

The model employed for the consensus filtering in this study is Gemini Flash 3 with the chain-of-thought parameter set to 'high'. These models, developed by Google, are one of the few models today with a very long context window of 1 million tokens allowing the insertion of a complete spreadsheet including Annexes A, E, and F of the IMDRF AET code list (1522 rows). Each prompt contains the instructions, a batch of 5 MAUDE adverse event descriptions, and the complete spreadsheet.

To account for the hierarchical structure of the IMDRF (A, E, and F) taxonomies, a hierarchical reconciliation protocol is implemented to identify consensus between human-annotated ground truth and LLM-generated codes. Con-

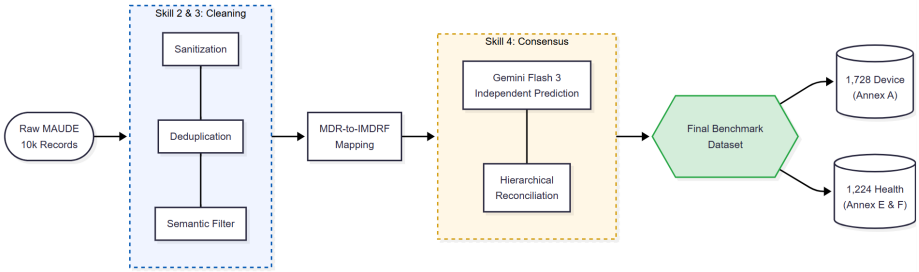

Figure 2: Pre-processing of MAUDE adverse event reports and their association with the IMDRF codes.

sensus is established at the IMDRF Level 2 Term granularity. Specifically, if two codes shared a lineage within the same Level 2 branch but differed at lower levels (Level 3), they are reconciled to their common Level 2 parent term. In cases where one code is a direct ancestor of the other, the most representative high-level term is retained to ensure semantic alignment. Pairs of codes belonging to distinct Level 2 branches are classified as non-consensus and excluded from the analysis. This approach ensures that the model's performance is evaluated based on its ability to identify the primary failure or clinical category, even if minor variations exist in technical sub-classification (at Level 3). Fig. 2 summarizes the steps of data cleaning, IMDRF mapping, and LLM-consensus.

The number of records left against the IMDRF device problem codes (Annex-A) reduces from 2899 to 1728 and against the IMDRF health effect codes (Annexes E and F) from 2899 to 1224 i.e., a 40% and 58% reduction in records, respectively. This constitutes the main ground-truth data against which the GPT-OSS-120B model is tested in Section 4. The proportion of discarded records (40–58%) merits comment in light of how MAUDE reports are written. As noted above, the majority of MAUDE filings are authored by manufacturers' regulatory affairs professionals under FDA 21 CFR Part 803 obligations [3], who routinely employ strategic, non-attributing language to avoid liability [9, 14]. A smaller fraction of discarded records is likely attributable to genuine human coding errors in the original MAUDE database, as documented in prior studies [14]. Retaining such records in an evaluation benchmark would introduce label noise rather than meaningful difficulty. The retained subset thus constitutes a *reliably codeable* layer of MAUDE reports where the narrative, despite strategic regulatory language, was clear enough for independent convergence between a human annotator and a state-of-the-art LLM — making it a well-defined dataset for evaluating the NLP workflows. Given the abundance of MAUDE data (millions of records available), the volume lost is operationally bearable against the benefit of a clean, low-noise ground truth.

3.2 Agent Skills for the Data Pre-processing – Justification via ISO 5338:2023

The pre-processing steps defined in Section 3 are defined with agent skills so that they can be reproduced by an NLP workflow. One example of such skills is defined in the Appendix at the end of this paper. Technically, each Agent Skill is implemented as a structured Markdown file (`skills.md`) containing: (a) a natural-language *objective* statement; and (b) an ordered list of *rule-based procedural steps* (e.g., regex patterns for character sanitization). This ensures that the reproducible, auditable steps are executed identically across runs, while LLM-dependent steps are isolated, versioned, and replaceable. To meet the engineering rigor required by ISO/IEC 5338, the skills-based approach moves the pre-processing from prompt experimenting to being part of a standardized life cycle process.

1. ISO 5338 recognizes that AI systems are heavily reliant on data and requires careful documentation of data provenance. By wrapping pre-processing steps as the skills, we can create a versioned audit trail. Every transformation – from the removal of non-alphanumeric characters to hierarchical reconciliation – is logged, allowing researchers to trace any output back to its original raw CSV file.

2. The standard treats AI model engineering as an extension of traditional software implementation (ISO 12207) [11]. Because the skills are based on a file-system with persistent logic, they mitigate the probabilistic drift often found in LLM-only data cleaning. This ensures that the workflow remains repeatable regardless of the specific model instance used, satisfying the reproducibility requirement of the standard.

3. ISO 5338 also emphasizes continuous validation. The skill architecture makes it easier to test the pre-processing logic independently. For example, we can verify the hierarchical reconciliation skill against a known subset of ground-truth data without needing to run the entire NLP workflow.

4 GPT-OSS-120B Testing

GPT-OSS-120B [1] is an open-source decoder-only model from OpenAI featuring 120 billion parameters and 131,072 tokens that is optimized for long-context understanding and reasoning. Its open-source transparent architecture, a sweet spot between 70B (more nuanced models) and widely-used models with world knowledge (400B+), and the ability to run on a single NVIDIA H100 (80GB) using MXFP4 quantization makes it a reasonable choice to be tested first against the dataset. Its context window is sufficient to perform a single long-context prompt containing the spreadsheet of Annexes A, E, and F of the IMDRF codes, a batch of 5 MAUDE adverse event descriptions, and the instructions. Unlike the NLP-workflows developed for ICD-10 coding, this single long-context

Table 1: Results of zero-shot classification of IMDRF codes by GPT-OSS-120B on the ground-truth adverse event reports across different granularity levels

Dataset	**Granularity**	**Accuracy**	**Precision**	**Recall**	F_1**-Score**
Device Problem	Level 1	0.5949	0.5368	0.5386	0.5377
	Level 2	0.5231	0.4505	0.4635	0.4569
	Level 3	0.3611	0.3022	0.3126	0.3073
Patient Problem	Level 1	0.6879	0.5214	0.6087	0.5617
	Level 2	0.5547	0.3616	0.4749	0.4106
	Level 3	0.2304	0.1274	0.1697	0.1455

prompt is tested because of the ease of use and readiness in practical scenarios because the list of codes to choose from is small (i.e., 1522). The model is allowed to predict multiple codes if it sees more than one suitable code for an adverse event.

Because each MAUDE record may carry more than one ground-truth IMDRF code and the model is permitted to predict multiple codes per record, the evaluation follows a multi-label classification protocol. Evaluation metrics are the accuracy, precision, recall, and F_1-scores derived from confusion-matrix components aggregated with averaging across samples. These evaluation metrics are calculated across three hierarchical levels of granularity: Level 1 (lenient), Level 2 (moderate), and Level 3 (stringent). The classification accuracy is defined as follows:

- Level 1 (Lenient): The prediction is considered correct if the Level 1 term matches the ground truth, irrespective of similarities within the secondary (Level 2) or tertiary (Level 3) hierarchical levels.
- Level 2 (Moderate): The prediction is considered correct if the match extends through the second hierarchical tier, effectively allowing for mismatches only at the third level.
- Level 3 (Stringent): The prediction is considered correct only when an exact match is achieved across all three hierarchical levels between the predicted and ground-truth codes.

An entry/row is considered correctly classified if at least one predicted code matches any ground-truth code at the level and this counts towards the overall Accuracy metric. Explicitly,

- True Positive is when the number of predicted codes match at least one ground-truth code.
- False Positive is when the predicted codes don't match any ground-truth code.

- False Negative is when the ground-truth codes aren't matched by any predicted code.

Table 1 summarizes the performance metrics across both problem categories and all three evaluation levels. The experimental results highlight a significant challenge in achieving cross-model alignment and accuracy when dealing with unstructured, long-context medical event descriptions. While the GPT-OSS-120B model provides relevant insights, it demonstrates an 'ill-disciplined' approach in picking up the common context defined by the human-Gemini consensus. Specifically, at Level 1 granularity, the model aligns with the consensus only 59.49% of the time for Device Problems and 68.79% for Patient Problems. This misalignment suggests that long-context inputs lead to divergent interpretations across different LLM architectures and that long-context window does not guarantee that the model will align with specific regulatory terminologies or maintain accuracy in zero-shot settings.

The model is performing equally at Level-1 for both the device problem and patient problem codes. However, the accuracy declines significantly for the patient problem codes at Level 3 in comparison to the device problem codes. This suggests that while the model is relatively capable of identifying that a patient was harmed (broadly), it struggles significantly more with the specific clinical "rigid specificities" of health effect coding than it does with mechanical device issues.

The data serves as a compelling benchmark for evaluating LLMs at Level 2 granularity. The results show that the GPT-OSS-120B frequently fails to align with Level 2 consensus terms, even when event descriptions are largely unambiguous. Because the ground-truth is itself a product of human-LLM consensus (requiring agreement up to Level 2), these gaps indicate that this open-source model may lack the contextual discipline required to navigate complex hierarchical codes consistently with well-defined expert baselines.

The discrepancy highlights that the scale of 120B parameters and a context window of 131,072 tokens (allowing all 1522 codes asked in the same prompt) does not guarantee alignment with specific regulatory terminologies. The model's divergence from the consensus at deeper levels (as low as 23.04% at Level 3) reflects a fundamental difficulty in reconciling diverse textual inputs with the rigid specificities of IMDRF coding. This variability is clearly noticeable in the long-context nature of the MAUDE database reports used for this study. This warrants future work on developing better NLP-workflows (despite the small list of IMDRF codes), just like the ones adopted for the much larger ICD-10 coding that have 68,000+ codes of diseases (e.g., [2, 11, 15]).

5 Conclusion

While AI agents offer transformative potential for medical device adverse event coding, this study advocates for deploying them after rigorous engineering standards. This paper argues for the adoption of agent skills to process a standardized benchmark dataset, following an ISO 5338-compliant pre-processing

methodology, hence establishing a foundation for systematic advancement in automated adverse event coding. The zero-shot GPT-OSS-120B results demonstrate the considerable technical challenges remaining before such systems can reliably support regulatory decision-making in medical device safety surveillance.

The discrepancy between open-source model performance and human-Gemini agreement-based ground truth suggests several avenues for further research: Despite a much smaller set of codes in comparison to ICD-10, this study warrants the use of specialized NLP-workflows for better zero-shot generalized classification of the IMDRF codes. This includes combining NLP models with ontology-driven reasoning systems and/or semantic searching to reduce the effective search space in the LLM chain-of-thought reasoning.

Limitations and Future Work. The benchmark covers the *reliably codeable* stratum of MAUDE — reports where manufacturer-driven strategic language did not prevent convergent coding by a human and an LLM. A meaningful future direction is developing NLP workflows that can also handle the strategically minimised narratives currently filtered out, as these constitute a substantial fraction of real-world manufacturer filings [9,14]. Additionally, the LLM-agreement filter relies on a single model (Gemini Flash 3); future work should replicate filtering with multiple LLMs and report inter-model agreement to produce a more robustly curated benchmark. Finally, future benchmarking should include multiple LLM architectures, especially domain-adapted clinical models, across varied prompting strategies (zero-shot, few-shot, retrieval-augmented generation) to provide a comprehensive evaluation picture.

References

[1] Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., Arbus, E.e.a.: gpt-oss-120b & gpt-oss-20b Model Card (Aug 2025). https://doi.org/10.48550/arXiv.2508.10925, http://arxiv.org/abs/2508.10925, arXiv:2508.10925 [cs]

[2] Boyle, J.S., Kascenas, A., Lok, P., Liakata, M., O'Neil, A.Q.: Automated clinical coding using off-the-shelf large language models

[3] Code of Federal Regulations: Part 803–Medical Device Reporting (2024), https://www.ecfr.gov/current/title-21/part-803

[4] International Medical Device Regulators Forum: Terminologies for Categorized Adverse Event Reporting (AER): terms, terminology and codes. Tech. Rep. IMDRF/AE WG/N43FINAL:2025, Geneva, Switzerland (Mar 2025), https://www.imdrf.org/documents/terminologies-categorized-adverse-event-reporting-aer-terms-terminology-and-codes

[5] International Organization for Standardization: ISO/IEC/IEEE 12207:2017 Systems and software engineering — Software life cycle processes. Tech. Rep. ISO/IEC/IEEE 12207:2017, International

Organization for Standardization, Geneva, Switzerland (2017), https://www.iso.org/standard/63712.html

[6] International Organization for Standardization: ISO/IEC 5338:2023 Information technology — Artificial intelligence — AI system life cycle processes (2023)

[7] International Organization for Standardization: ISO/IEC JTC 1/SC 42 information technology — Artificial intelligence — AI system life cycle processes (2023), https://www.iso.org/standard/81118.html

[8] Kass-Hout, T.A., Xu, Z., Mohebbi, M., Nelsen, H., Baker, A., Levine, J., Johanson, E., Bright, R.A.: OpenFDA: an innovative platform providing access to a wealth of FDA's publicly available data. J. Amer. Med. Inform. Assoc. **23**(3), 596–600 (May 2016). https://doi.org/10.1093/jamia/ocv153, https://academic.oup.com/jamia/article/23/3/596/2908999

[9] Knisely, B.M., Levine, C., Kharod, K.C., Vaughn-Cooke, M.: An Analysis of FDA Adverse Event Reporting Data for Trends in Medical Device Use Error. Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care **9**(1), 130–134 (2020). https://doi.org/10.1177/2327857920091024, https://doi.org/10.1177/2327857920091024, _eprint: https://doi.org/10.1177/2327857920091024

[10] Li, S., Zheng, C., Wu, J., Xu, Q., Xu, X., Wang, H., Sun, Y., Bai, Z., Xu, Y., Zhu, L., Hu, W., Huang, F.: Verification is All You Need: Prompting Large Language Models for Zero-Shot Clinical Coding. IEEE J. Biomed. Health Inform. **29**(11), 8536–8549 (Nov 2025). https://doi.org/10.1109/JBHI.2025.3593028, https://ieeexplore.ieee.org/document/11097877/

[11] Li, X., Feng, Y., Gong, Y., Chen, Y.: Assessing the Reproducibility of Research Based on the Food and Drug Administration Manufacturer and User Facility Device Experience Data. J. Patient Saf. **20**(5), e45–e58 (Aug 2024). https://doi.org/10.1097/PTS.0000000000001220

[12] Luschi, A., Nesi, P., Iadanza, E.: Evidence-based clinical engineering: Health information technology adverse events identification and classification with natural language processing. Heliyon **9**(11) (Nov 2023). https://doi.org/10.1016/j.heliyon.2023.e21723

[13] MedTech Europe: International nomenclature implementation in European incident reporting with medical devices. Tech. rep. (May 2020), https://www.medtecheurope.org/2020/05/06/international-nomenclature-implementation-in-european-incident-reporting-with-medical-devices/

[14] Mishali, M., Sheffer, N., Mishali, O., Negev, M.: Understanding Variation Among Medical Device Reporting Sources:

A Study of the MAUDE Database. Clin. Ther. **47**(1), 76–81 (Jan 2025). https://doi.org/10.1016/j.clinthera.2024.10.004, https://linkinghub.elsevier.com/retrieve/pii/S0149291824002947

[15] Mustafa, A., Naseem, U., Rahimi Azghadi, M.: Can reasoning LLMs enhance clinical document classification? Health Technol. (Jan 2026). https://doi.org/10.1007/s12553-025-01041-y, https://link.springer.com/10.1007/s12553-025-01041-y

[16] Shi, Y., Yang, E., Gong, Y.: Unveiling Endoscopic Procedures in MAUDE: An ETL-LLM Method. In: Mantas, J., Hasman, A., Gallos, P., Zoulias, E., Karitis, K. (eds.) Studies in Health Technology and Informatics. IOS Press (Jun 2025). https://doi.org/10.3233/SHTI250680, https://ebooks.iospress.nl/doi/10.3233/SHTI250680

[17] Uciteli, A., Kropf, S., Weiland, T., Meese, S., Graef, K., Rohrer, S.: Ontology-based specification and generation of search queries for post-market surveillance. J Biomed Semant **10**(1), 9 (Dec 2019). https://doi.org/10.1186/s13326-019-0203-7, https://jbiomedsem.biomedcentral.com/articles/10.1186/s13326-019-0203-7

[18] U.S. Center for Disease Control and Prevention: Health Insurance Portability and Accountability Act of 1996 (1996)

[19] U.S. Food and Drug Administration: Manufacturer and User Facility Device Experience (MAUDE) Database. Tech. rep., https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfmaude/search.cfm

[20] Wang, E., Kang, H., Gong, Y.: Generating a health information technology event database from FDA maude reports. In: Studies in Health Technology and Informatics. vol. 264, pp. 883–887. IOS Press (Aug 2019). https://doi.org/10.3233/SHTI190350

[21] Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A., White, R.W., Burger, D., Wang, C.: AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversations (2024)

[22] Wunnava, S., Harris, D., Bourgeois, F.T., Miller, T.A.: Development of a Benchmark Corpus for Medical Device Adverse Event Detection. Tech. rep. (May 2024)

Appendix

This appendix provides an example of how one can use the MAUDE preprocessing skills in one's automatic NLP-workflow with the agent skills (saved in files like skills.md). The writing of these skills are generalized for the paper and can be made more specific given the file architecture of one's project.

These skills text are generated by Gemini Flash 3.0 and edited + reviewed by the authors.

Skill 1: Data Acquisition and Selection

- **Objective**
- Defined steps: Standardize the retrieval of raw adverse event records.
    1. **Query Parameters**: Define specific search parameters (e.g., date range 2016–2025, case types: Death, Malfunction, Serious Injury).
    2. **Source Verification**: Validate the API endpoint or CSV source (OpenFDA or MAUDE manual export) to ensure data provenance.
    3. **Privacy Guardrail**: Confirm the absence of PII (Personally Identifiable Information) at the point of ingestion to comply with HIPAA/FDA guidelines.

Skill 2: Data Integrity & Cleaning

- **Objective**: Normalize the raw CSV input for NLP consumption.
- Defined steps:
    1. **Redundancy Check**: Identify and remove duplicate entries within and across the CSV file.
    2. **Character Sanitization**: Execute a regex-based routine to strip all non-alphanumeric and non-punctuation characters from event descriptions to prevent tokenization errors.
    3. **Completeness Validation**: Filter out any record missing its corresponding Device Problem or Health Effect codes.

Skill 3: Semantic Filtering & Noise Reduction

- **Objective**: Eliminate uninformative narratives that lack clinical utility.
- Defined steps:
    1. **Exclusion Matching**: Use semantic matching to identify and remove descriptions coded as "Insufficient Information," "Adverse Event Without Identified Device/Use Problem," or "Appropriate Term/Code Not Available".
    2. **MDR-to-IMDRF Translation**: Map the FDA's Medical Device Report (MDR) codes to the standardized IMDRF (International Medical Device Regulators Forum) AET codes using the provided mapping files.

Skill 4: Hierarchical Reconciliation (The “Consensus" Skill)

- **Objective**: Ensure ground-truth alignment at the IMDRF Level 2 granularity.
- Defined steps:
  1. **Lineage Identification**: Analyse the hierarchical structure of IMDRF Annexes A, E, and F.
  2. **Branch Reconciliation**: If two codes share a Level 2 parent branch, reconcile them to that common ancestor; otherwise, classify as non-consensus and exclude.
  3. **Audit Log Generation**: Record the reduction in records (e.g., the 40% reduction in Annex-A and 58% in Annexes E/F) to document data lineage.