Scientific Reports

https://doi.org/10.1038/s41598-026-43961-3

Article in Press

Classification of health product defect reports by deep learning

Received: 20 January 2025

Accepted: 9 March 2026

Published online: 14 March 2026

Cite this article as: Sancenon V., Huang Y., Zou L. *et al.* Classification of health product defect reports by deep learning. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-43961-3

Vicente Sancenon, Yiting Huang, Lin Zou, Desmond C. H. Teo, Sreemanee R. Dorajoo, Pei San Ang, Han Leong Goh & Andy W. A. Ta

Classification of health product defect reports by deep learning

Vicente Sancenon[1,3,*], Yiting Huang[1,3], Lin Zou[1], Desmond C. H. Teo[2], Sreemanee R. Dorajoo[2], Pei San Ang[2], Han Leong Goh[1], Andy W. A. Ta[1]

[1] Synapxe, 6 Serangoon North Ave 5, #01-01/02, Singapore 554910, Singapore. [2] Health Sciences Authority (HSA), 11 Biopolis Way, #11-01 Helios, Singapore 138667, Singapore. [3] These authors contributed equally: Vicente Sancenon, Yiting Huang.

*email: vicente.enrique@synapxe.sg

Abstract

Quality defects in substandard medicines represent a threat to public health. Rapid and accurate identification of these defects is critical to prioritise the cases and implement regulatory measures. However, the development of surveillance systems to classify reports of health product defects remains a largely unmet need in medicine safety monitoring. The objective of this study is to implement an AI system to support the classification and prioritization of health product defect reports. To develop a deep learning system for the classification of health product defect reports, 13,830 reports collected between 2010 and 2021 were used. The reports were labelled by a panel of pharmacovigilance experts into 21 categories following standardised medical terminology. Our system harnesses state-of-the-art language algorithms that extract rich textual features to classify the reports. The functionality of the system is enhanced with explainable features that provide interpretability and actionable insights to decision-makers. Our system achieves top-1, top-2, and top-3 accuracies of 86%, 93%, and 96%, respectively. There is a statistically significant positive correlation between sample size and top-1 (Pearson's r: 0.643; 95% CI: [0.2921, 0.8411]; p-value: 0.0016), top-2 (Pearson's r: 0.735; 95% CI: [0.4439, 0.8856]; p-value: 0.0001), and top-3 (Pearson's r: 0.635; 95% CI: [0.2808, 0.8374]; p-value: 0.0020) performance metrics. Likewise, model accuracy is positively correlated with confidence scores (Pearson's r: 0.927; 95% CI: [0.8253, 0.9703]; p-value <0.00001). A feature analysis reveals that the most influential words in model decision are conceptually and semantically related to their respective product defect categories. Our model has been validated with prospective data. The developed classification system allows for standardisation in case triage, and potentially improves case prioritisation and processing workflows, leading to more prompt response for quality defects with high public health impact.

Keywords

Deep Learning, Drug Safety, Health Information Management, Health Product Defects, Large Language Model (LLM), Medical Dictionary for Regulatory Activities (MedDRA)

Introduction

Post-marketing defects in health products represent potential threats to public health due to lack of quality, safety, and/or efficacy[1]. The magnitude and impact of such defects is illustrated by the fact that in fiscal year 2022, the FDA classified 343 recall events relating to more than 1,500 violative drug products, including 64 that were potentially life-threatening[2]. Timely understanding and diagnosis of the nature of issue and root cause of defects, especially those that can potentially lead to adverse events, can guide effective regulatory oversight, enabling more risk-appropriate corrective and preventive measures to be taken, and hence safeguarding public health and safety. In Singapore, the Health Sciences Authority (HSA) is the national authority that regulates and monitors the post-market safety of health products to safeguard public health. Besides monitoring the spontaneous adverse event reports from healthcare professionals, the HSA Vigilance and Compliance Branch routinely receives product defect reports associated with substandard medicines from multiple sources and assigns them a priority level based on the nature of the defect and the potential risk to public health. To classify the nature of product defects and facilitate priority review, HSA has adapted Medical Dictionary for Regulatory Activities (MedDRA) terminology, a standardised collection of medical terms created by the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) to facilitate the reporting of health product defect issues and used by health regulators internationally to exchange information[3]. HSA has developed a customised list of MedDRA terms that encompass the various types of product defect reports received, denoted as MedDRA-HSA[4]. MedDRA-HSA terms are hierarchically structured into Highest Level (HLT), Preferred (PT), and Lowest Level (LLT) terms. LLT provide the finest granularity and therefore the highest specificity to categorise a product quality issue. There are a total of 30 MedDRA-HSA LLT terms grouped into high, medium, and low severity categories based on their estimated impact to public health (Supplementary Table S1). Identifying the correct MedDRA-HSA term is a critical decision-making checkpoint as it determines the priority of each case for subsequent intervention[4]. Currently, HSA vigilance officers perform manual review of reports to determine the nature of the defects. This process is laborious and time-consuming due to both complexity and workload. There is also potential for inconsistency in classifying the reports given the subjectivity in case assessment, which may delay the detection of high-priority cases and have negative downstream consequences.

Artificial intelligence (AI) is emerging as a powerful decision support tool to enhance clinical workflows[5]. In recent years, advances in NLP research have led to significant progress in natural language understanding (NLU) and generation (NLG), largely enabled by the rise of a novel attention-based neural network, the transformer[6]. Transformers bring two major advantages over previous architectures: 1) Computational efficiency; and 2) Ability to capture long distance dependencies[7–9]. Transformers surpass traditional Recurrent Neural Networks (RNNs) and probabilistic models achieving state-of-the-art (SOTA) performance in multiple NLP tasks[6,10–18]. Since their inception, a progressive escalation in transformers size has

led to the emergence of large language models (LLMs) or foundation models[19,20]. Typically, LLMs are pre-trained with vast amounts of text corpora[12,13,16] and subsequently fine-tuned with domain-specific data[21,22]. One of the most successful and broadly studied foundation models is Bidirectional Encoder Representations from Transformers (BERT)[23]. BERT was originally pre-trained on two NLP tasks: Masked Language Modeling (MLM), where it predicts hidden words in a sentence, and Next Sentence Prediction (NSP), where it determines if two sentences follow each other in the original text, enabling it to learn deep contextual understanding for various NLP tasks. BERT has been thereafter extensively fine-tuned by researchers in the biomedical, communication, acoustics, linguistics, education, and legal domains[24–38]. Among other innovations, BERT introduced the special token [CLS], which captures global feature representations of documents and can be used for text classification. The base configuration of BERT comprises 12 attention layers and 110M parameters[23]. An advantage of BERT base over other foundation models is its relatively lightweight size and high performance[39–41].

Despite its proven success, the fine-tuning paradigm requires considerable compute power and storage capacity. More recent strategies leverage on conditioning a pre-trained frozen model to perform a task by providing a few demonstrations (few shot learning)[42] or by reformulating the task as a cloze question (zero shot learning)[43,44]. In the past these frameworks have been successfully applied to LLMs[12,13,16] and more recently have been adapted to smaller LMs[45,46]. In context conditioning, a short sequence of tokens, or prompt, is prepended to the input sequence to query a pre-trained model with the aim of extracting factual knowledge[47]. Prompt embeddings are learnt during training while keeping the rest of the model parameters frozen[48,49]. Furthermore, prompts can be prepended to all the layers of a transformer model (deep prompts)[50,51]. Notably, deep prompt tuning has been shown to achieve comparable results to fine-tuning across NLP tasks and model scales[52].

The purpose of this study was to develop and characterise an AI system to classify reports of health product quality defects using MedDRA-HSA terminology. The system is designed as a decision support tool, rather than a fully automated system, to reduce manual labour, assist officers with case prioritization, and accelerate regulatory intervention to protect public health. We selected a BERT-based architecture due to its lightweight and ability to extract long term syntactic and semantic dependencies[7–9] compared to less efficient RNNs[18,53]. Related work using AI to predict MedDRA terms focused on the identification of adverse drug reactions (ADRs) rather than medical product defects. One study developed a rule-based scoring system to classify ADRs based on word matching[54]. This system is only suitable for very short descriptions of up to 20 characters and loses accuracy for longer narratives. In addition, rule-based systems fail to capture contextual information, word derivations, and do not adapt to changes in data patterns. Two more recent studies adopted MedDRA terminology to identify ADRs from drug labels[55] or patients' reports[56] using machine learning and deep learning. These methods achieve higher performance than rule-based algorithms. However, the machine learning methods rely on sparse term frequency features that omit richer

contextual information, while the deep learning architectures chosen (CNNs and RNNs) are inferior to transformers in retaining long range dependencies[18].

While previous work to identify ADRs showed promising results, automated classification of product defect reports following MedDRA terminology remains an unmet need for drug regulators. In contrast to previous literature, our study implemented an AI system to assists drug regulators in classifying the nature of health product quality issues in documents using text-based case descriptions. In addition, this work brings 3 additional advances: 1) The implementation of an efficient transformer-based architecture; 2) The enhancement of fine-tuning performance with prompt-based learning; 3) The provision of explainability features to assist officers in case review and decision-making. Our system achieves top-1 accuracy of 86% on the classification of the 21 most common categories of MedDRA-HSA product defects, which represent 99.2% of all cases. We show that the model can be custom-tuned to further increase its accuracy and recall. Due to its lightweight the model is easily transferable to domain-related tasks and deployable as a web application or API.

Results

Performance evaluation

Overall performance

In order to develop robust text classifiers that captures contextual information to classify medical product defect alerts into the 21 most common MedDRA-HSA terms, we implemented a deep learning model using a backbone transformer architecture (BERT base) as feature extractor and adapting it to a sequence classification task (Methods). We call our fine-tuned model MedDefects-BERT. MedDefects-BERT achieved a top-1 accuracy (across all 21 MedDRA-HSA terms) of 86%, a macro F1 score of 72%, and a macro average Recall of 72% in the test set (Table 1). MedDefects-BERT outperformed two baseline probabilistic classifiers that rely only on the distribution of MedDRA-HSA terms in the training dataset (Wilcoxon signed-rank test; p-value <0.00001) and on word frequency counts (Wilcoxon signed-rank test; p-value <0.00001) (Supplementary Table S2). Notably, MedDefects-BERT outperformed two BERT models specialised in the clinical domain [Bio_ClinicalBERT (Wilcoxon signed-rank test; p-value: 0.0385) and Bio_Discharge_Summary_BERT (Wilcoxon signed-rank test; p-value: 0.0434)] and matched the performance of one BERT model specialised in the scientific domain [SciBERT (Wilcoxon signed-rank test; p-value: 0.395)] fine-tuned with identical methodology (Supplementary Table S3), indicating no additional benefit of using specialist BERT models over BERT-base as the initial checkpoint for fine-tuning.

Positive and negative performance

To assess the effect of the variety and number of reports on the positive and negative performance of MedDefects-BERT during batch inference, we conducted two spiking experiments. In the experiments we randomly sampled subsets of 1) the 21 MedDRA-HSA terms, and 2) the 5 sources of product defect reports used in this study from the test dataset and subsequently injected them with positive or negative samples of each group. We then computed the accuracy score for each group pre and post-spiking (Methods). Supplementary Tables S4 and S5 show that there are no statistically significant differences in the performance of the classification system across the 21 classes of product defects before and after spiking the test data with positive samples or across the 5 sources of reports before and after spiking the test data with positive samples and negative samples. In addition, there is a statistically significant increase in model performance across the MedDRA-HSA terms (+6% average accuracy increment) after spiking the test data with negative samples. These findings indicate that the positive and negative performance of MedDefects-BERT is not affected by the diversity and quantity of MedDRA-HSA classes or report sources during batch inference.

Top-k multi-label performance

In health product defect prioritisation, missing high impact cases (type II errors) is more critical than overrating low impact ones (type I errors). In addition, product defects may not always belong to a unique MedDRA-HSA label, but suit multiple

categories. Therefore, in a decision support scenario, it may be more effective to suggest more than one defect category to the surveillance officer. Although this approach increases ambiguity, it minimises the chances of type II errors with little impact on the workflow. When considering the 3 most probable MedDRA-HSA terms predicted by the model, top-3 accuracy increased to 96%, macro F1 score to 92%, and macro average Recall to 91% (Table 1).

At a lower granularity level, the model identified “Product adulterated and/or contains prohibited substance” with the highest top-1 F1 score (95%) and “Product label/leaflet issues NOT impacting strength, dose and/or safety” with the lowest (27%). This disparity is likely caused by the skewed distribution of samples across MedDRA-HSA terms in the development and test sets (Table 2). We found a statistically significant positive correlation between the sample size in the development set and the top-1 (Pearson’s r: 0.643; 95% CI: [0.2921, 0.8411]; p-value: 0.0016), top-2 (Pearson’s r: 0.735; 95% CI: [0.4439, 0.8856]; p-value: 0.0001), and top-3 (Pearson’s r: 0.635; 95% CI: [0.2808, 0.8374]; p-value: 0.0020) macro F1 scores achieved in the test set (Fig. 1a; Supplementary Table S6), suggesting that collection of more samples from the minority classes would presumably improve class-specific and overall model performance. Remarkably, top-3 predictions increased the F1 scores above 70% across all MedDRA categories irrespective of sample size, significantly improving the model robustness to correctly identify product defect reports.

Error analysis

An error analysis revealed 4 possible causes for 93.5% of the misclassified reports while the remaining 6.5% remain unclear (Table 3). The majority of errors (50.6%) may be attributed to the presence of keywords that are strongly associated with certain MedDRA terms other than the ground truth. Other possible sources of error include potential class label overlap (24.2%), multi-defect reports (10.0%), and incomplete defect description (8.7%). A common example of case misclassification is distinguishing between “Contamination with chemical substance” and “Product adulterated and/or contain prohibited substance”. “Adulteration" involves the deliberate, often illicit, modification of a product's content, whereas "contamination with chemical substance" is often the undesired presence of chemical substances in the product that were not intended to be present. For example, a vitamin & mineral supplement was reported to contain beta-phenyl-gamma-aminobutyric-acid. There was insufficient information at the initial receipt of the report to determine whether the contamination was deliberate or unintended. In this case, our tool predicted the nature of issue label to be "Product adulterated and/or contains prohibited substance", where in fact the correct label was “Contamination with chemical substance” as it was not an illicit activity. More details on the various categories of misclassified cases and some common examples are available in Table 3. Collectively, this analysis supports the multilabel view, where multiple rather than a unique MedDRA term may be assigned to each report.

Thresholding analyses

Classification models output a confidence score associated to each class label as a probability distribution over all classes. Fig. 1b shows the median confidence score of the most probable MedDRA-HSA label predicted by the model as a function of the F1 score across the 21 labels. In general, MedDRA-HSA terms with higher rate of misclassification errors (lower F1 score) were predicted with lower confidence scores and vice-versa. This correlation is statistically significant (Pearson's r: 0.927; 95% CI: [0.8253, 0.9703]; p-value <0.00001) (Supplementary Table S6), indicating that higher confidence predictions are more likely to be correct. Fig. 1c, shows the distribution of the confidence scores for each class label on the test set for correctly classified and misclassified samples. In general, misclassified cases were predicted with lower confidence than correctly classified cases in the 18 MedDRA categories with the highest F1 scores. This observation suggest that confidence thresholds may be effective to augment prediction accuracy.

Medical product defects can be grouped into multiple levels of severity based on their potential impact to public health (Table 2). When deploying an AI model for decision support in regulatory settings, it is crucial to maximise the recall of high impact cases. This objective can be attained by setting lower bound probability thresholds to the model confidence scores. Predictions with a probability below the threshold are annotated as "uncertain" whereas the rest ("certain") are annotated with the predicted product defect class. This strategy increases prediction accuracy at the cost of increasing uncertainty. To ascertain the effect of increasing confidence thresholds on the model accuracy and recall, we performed a threshold analysis. Fig. 1d shows the top-1 accuracy, macro F1-score, Precision, Recall (of both total and above-threshold predictions), and the fraction of uncertain cases as a function of confidence threshold. A threshold of 0.78 increased top-1 accuracy to 91%, macro average F1 score to 76%, and macro average Recall to 76% while only returning 14.3% of all cases as uncertain (Table 1), thus maximising the recall over the certain cases and keeping the uncertainty rate low. This parameter can be tuned to attain a tolerable trade-off between performance and uncertainty.

Performance validation with prospective data

Our system development is retrospective in nature, since only historical records were used for both training and validation. As such, any model developed using historical data and implemented for future predictions is at risk of data and concept drifts. To measure the potential impact of data and/or concept drifts in the performance of our system, we tested the performance of MedDefects-BERT with a dataset of reports collected prospectively. MedDefects-BERT showed comparable accuracy on the prospective and retrospective datasets, with no significant differences in metrics across product defect categories (Supplementary Table S7). This finding indicates that our model is robust to potential changes in data distribution or that no significant drifts occurred in the prospective data.

Interpretability analysis

A) Global feature analysis

Deep learning models extract dense features that capture rich and complex dependencies between text entities[7–9] and thus achieve higher performance in NLU tasks than bag-of-words (BOW) and other probabilistic models, which only capture statistical features. In particular, BERT models represent documents as 768-feature vectors that can be used for downstream NLP tasks, such as classification[23]. In order to visualise the features extracted by MedDefects-BERT, we projected the multidimensional representations of 8 classes of documents from the development set into a lower 2-dimensional space using the Uniform Manifold Approximation and Projection (UMAP) algorithm[57] (Methods). Fig. 2 shows that the 8 categories of product defects cluster in distinct regions of the vector space, confirming that MedDefects-BERT extracts distinct features that are suitable to discriminate between unrelated MedDRA terms. In addition, misclassified documents are projected outside the boundaries of their ground truth cluster. These findings provide insights into the model misclassification errors and can potentially be helpful to further improve model accuracy, for example by refining the pre-processing pipeline to obtain more consistent vector representations of documents within each class.

B) Local feature analysis

The UMAP analysis provides an explanation for the misclassification based on global (document-level) features extracted by the model but does not take into consideration local (i.e. word-level) features. Moreover, dense vector representations of documents are not directly interpretable, and consequently actionable, by humans. To investigate the importance of individual tokens to the model outputs, we conducted a SHapley Additive exPlanations (SHAP) analysis[58] of MedDefects-BERT (Methods). Table 4 displays a global summary of the top-10 tokens with the highest average Shapley values in each class. Overall, the most important tokens for each class appear to be conceptually and semantically correlated to their respective MedDRA terms. For example, the words ‘fungi’, ‘transmission’, and ‘variant’ are among the top 10 most influential terms to classify a product defect report as ‘Contamination with microbes’ (High severity), the words ‘sediment’, ‘precipitation’, and ‘deposit’ as ‘Product deposit’ (Medium severity), and the words ‘shaved’, ‘squash’, and ‘oversized’ as ‘Product physical issue’ (Low severity). This analysis provides a high-level explanation for the model decision-making factors.

In addition to the global summary, SHAP analysis also provides instance-level explanations of the model predictions that can be easily verified and actioned by domain experts. The analysis highlights the importance of each token in a particular document by a colour gradient and displays a force plot that ranks the contributions of all tokens to the final classification output. Fig. 3 displays the SHAP analysis of a single instance in the test dataset. Consistent with the expectations of the transformer architecture, the SHAP analysis illustrates how MedDefects-BERT not only attends to individual tokens, as probabilistic models do, but also to richer contextual information. The provision of explainable features that are intelligible to humans can assist officers to quickly verify the model decisions. Therefore, this

explainability analysis provides the end-user with a friendly interface for human-machine interaction and a visualisation tool for rapid decision making.

Model improvement

With the aim of improving the performance of MedDefects-BERT, we further tuned the model to learn deep prompts, while freezing the rest of the model parameters. We call the resulting model MedDefects-DPT-BERT. Notably, this refinement increased the F1 score of the model in 12 out of 21 categories, while decreased the score in only 6 (Table 1). Our findings indicate that fine-tuning and prompt-tuning are not mutually exclusive strategies but can used in combination to synergise outcomes of small language models.

Discussion

The ISO Identification of Medicinal Product (IDMP) and the Global Substance Registration System (GSRS) are extraordinary resources developed to help identify and accommodate all of the substances found in global commerce. The goal of such projects is to help facilitate the understanding of the relationships between substances and products from a quality, safety, and drug utilization perspective throughout the world. This knowledge can be useful to help identify and avoid manufacturing defects from re-occurring around the world[59]. In this study, we have developed and validated a deep learning model to identify the 21 most common classes of health product defects in company reports and regulatory alerts following standardised MedDRA terminology.

The observed correlation between sample size and F1 score helps to elucidate the minimum number of cases required to develop a model with a target performance. For example, Fig. 1a shows that if the target was to achieve a top-1 F1 score of 80% or above in a particular MedDRA-HSA category, the required number of samples would be around 1000, whereas for a top-3 equivalent performance, a relatively small set of 50 reports would suffice. These insights can guide future data collection efforts by prioritising the most relevant class labels where misclassification errors are more critical.

The strong correlation between confidence and F1 score enables to gauge the reliability in individual model predictions and to set up acceptability criteria for further scrutiny by human experts. For example, Fig. 1b shows that, on average, to achieve a top-1 F1 score of 80% or above, confidence scores should be around 0.90. The thresholding analysis in Fig. 1d supports this claim: Rising the confidence threshold increases the accuracy of the model outputs at the cost of growing uncertainty, i.e., the number of predictions that do not pass the confidence acceptability criteria, and therefore requires deeper human intervention. These findings inform on the selection of adequate cut-off values to meet desired reliability targets with an acceptable level of incertitude.

Related studies have developed models to identify ADRs for pharmacovigilance use-cases[54–56,60]. The scope of these studies is different from the present one as they perform token-level classification tasks to extract the segments of text that contain mentions to ADRs. In contrast, the goal of our study was to develop a document-level classifier of product defect reports using MedDRA terminology. From a technical perspective, previous studies focused on rule-based, regex, or machine learning algorithms. Those algorithms have the limitations of being restricted to human defined rules or to simple text representations, such as term frequency features. Subsequent studies implemented more sophisticated RNNs to capture richer contextual features. However, the performance of those models degrades with sequence length due to the inability of those features to retain long distance dependencies[18]. By contrast, our study took advantage of the more optimised transformer architecture, which extracts far reaching dependencies via an attention mechanism and is computationally more efficient, delivering SOTA results across NLP tasks and domains[24–38]. By adapting the transformer architecture to a

classification task and applying transfer learning[61], we could fine-tune our model efficiently.

To better understand the underlying reasons behind the model correctness and errors, we investigated the properties of the feature vectors extracted by the transformer backbone. A document-level analysis revealed that correct and erroneous predictions can be explained by the topology of the corresponding vector representations in a multidimensional space. While correctly classified reports within the same MedDRA label tend to cluster together, misclassified ones deviate significantly from their expected regions. This analysis indicates that for misclassified documents, the model failed to extract features which are congruent to other documents within the same label. This phenomenon may be attributed to three main (not mutually exclusive) reasons. First, the contents of the misclassified document are substantially different from those of other documents from the same category, making it an intra-class outlier. Second, the sample size of the MedDRA label to which the misclassified document belongs is not sufficiently large for the model to learn features that can be generalised to the entire class. Limited training data can lead to biased classifications, whilst evolving language patterns and domain-specific terminology may not be adequately represented in the model's knowledge base. Third, the model fails to properly extract features from the misclassified document due to intrinsic limitations, such as input size or vocabulary repertoire. Furthermore, models may struggle with nuanced language, context-dependent meanings, or ambiguous phrasings that humans could interpret accurately with sufficient context and experience. These factors may result in content being incorrectly classified. The first two limitations may be addressed by collecting a wider variety of samples from the minority classes, thus enabling the model to learn more generalizable embeddings during training. The third limitation may be addressed by further improving the processing pipeline to extract more consistent vector representations of documents within each class.

Deep learning model predictions are complex to interpret solely in terms of the corresponding feature outputs. In decision support set-ups, it is generally more useful to understand model predictions in terms of the actual input data, rather than the intermediate or the final feature maps. Our system provides a user-friendly explainability tool that highlights the individual tokens and segments the input text with a colour gradient according to their positive or negative influence in the final model decision. These explanations can be easily contrasted by domain experts to verify the validity of the model predictions. This tool adds interpretability and insights for the officers to make more expedite and informed decisions. Indeed, a global analysis revealed that the most important tokens for the classification of the 21 MedDRA-HSA terms considered in this study are semantically and conceptually associated with their corresponding class labels. This finding reinforces the notion that during fine-tuning the model learns word embedding representations that are both individually and contextually suitable for the task of classifying product defect reports.

Fine-tuning is currently the most widely used approach to adapt pre-trained language models to domain-specific tasks. However, more recent strategies rely on

conditioning frozen language models with a prompt to perform a task[13,44]. One such strategy is prompt tuning, which consists in learning the optimal prompt embeddings for a specific task while keeping the rest of the model parameters frozen[48,49]. An improved variation of this methodology is deep prompt tuning, where a prefix is prepended to all model layers and not just the input layer[50–52]. Empirical evidence shows that prompt tuning and deep prompt tuning achieve comparable performance to fine-tuning for moderate to large language models (330M parameters and above) in multiple NLP tasks[52,62]. However, the effectiveness of prompt tuning for smaller language models such as BERT base (110M parameters) and complex multiclass classification tasks (> 20 classes) has not been tested experimentally. Although prompt-based learning is typically used as an alternative to fine-tuning, we show that, in combination, both strategies can synergise to deliver superior outcomes. In our case, model performance was enhanced by prepending learnable soft prompts to all the model layers. Our MedDefects-DPT-BERT model surpassed the prediction accuracy of MedDefects-BERT in 12 out of 21 MedDRA categories and improved the macro average recall to 73%. Our experiments demonstrate that lightweight fine-tuned models, such as BERT base, can be successfully conditioned with deep prompts to achieve higher performance.

Despite its high accuracy, our model is susceptible to further revisions and improvements. First, despite their efficiency, transformers have a limitation to their input sequence size. In particular, BERT has a maximum context window of 512 tokens, including special tokens. Therefore, relevant information located in the last sections of long documents may be disregarded by the model. Although we applied some filters to extract key information from certain reporting templates, additional improvements could be implemented to the processing pipeline to generalise the extraction of critical information from all sources. Second, although the MedDRA terms in our dataset were currently reviewed by domain experts to add consistency to the standard reference labels, certain products defects may qualify to be annotated with multiple terms due to the complexity of the issues, which makes it challenging to evaluate the accuracy of the model. An error analysis confirmed that a substantial number of the misclassified reports 1) may indeed be reclassified with alternative, non-mutually exclusive, terms, as they describe multiple defects (10.0%), or 2) have a ground truth label that depends on subtle contextual information, not always available in the report (24.2%). To deal with this ambiguity, the classification task could be reformulated as a multilabel use-case. Third, our dataset is highly imbalanced. Our analyses show that per class performance is correlated with class size. This observation indicates that class-specific and overall model performance would benefit from gathering more reports from those categories. One possibility would be to collate data from more countries and jurisdictions to enrich the size and diversity of the training dataset, particularly of the minority classes. While this approach is limited to the availability of actual cases collected, other data augmentation methodologies, such as resampling strategies and generation of synthetic reports, are feasible. Future work will also need to address the generalisation of the model to the full set of 30 MedDRA-HSA terms, including those that were excluded from this study due to small sample sizes. For example, one strategy to circumvent the limitations imposed by the sample size of

the rare categories would be to implement a hierarchical system where multiple modules collaborate to sequentially predict the High, Preferred, and Lowest Level terms. This approach would reduce the number of categories predicted by each module and therefore minimise the class imbalance faced by the single-model approach. In combination with oversampling and undersampling techniques, the hierarchical approach would allow to incorporate rare categories into the classification system by reducing the number samples required to train the modules responsible for the classification of the minority classes. One additional strategy to improve the system's performance on minority classes would be to enrich the dataset with synthetic samples. While this strategy is technically feasible due to the emergence of generative AI, it holds concerns about data privacy protection and may be impractical in resource constrained environments.

Detection and management technologies of quality defect issues associated with substandard medicines have been significantly understudied in general, and yet they form an essential component of the entire ecosystem of medicines safety monitoring. In particular, substandard medicine issues can be wide-ranging with varying severity of impact to public health. Notably in recent years, defect issues that affected healthcare systems globally, such as the detection of nitrosamine impurities in medicines (https://www.hsa.gov.sg/announcements/press-release/hsa-recalls-three-brands-of-losartan-medicines-from-hetero-labs-ltd), or quality issues with COVID-19 vaccines that affected national vaccination programs to combat the COVID-19 pandemic[63], have put health product quality surveillance in the spotlight. Enhanced vigilance systems may thus be crucial to prevent the widespread development of such issues. Our study describes the development of a deep learning system, in particular a transformer architecture, to assist drug regulators in the task of classifying the nature of health product defects reported in text files, specifically using MedDRA-derived terminology. Our system can potentially enhance current workflows and augment the detection and management of health product defects cases in multiple ways. First, the system can assist vigilance officers in reviewing the reports, thus increasing productivity, facilitating standardisation of nature of defect issues, streamlining case prioritisation, and accelerating case management. Second, the system can be customised to attain desirable levels of recall with little impact on uncertainty or ambiguity. For example, the model can suggest multiple options to the officer, rather than a single one, including some that may not be immediately obvious, hence increasing the probability of a correct diagnostic. Custom confidence thresholds can be further adjusted to reach a tolerable trade-off between accuracy and uncertainty. Third, the system is further enhanced with additional explainability tools that facilitate decision making by human experts. These tools highlight the segments of text that are most influential in the prediction, thus adding interpretability to the model decisions and actionable insights for decision-making. Finally, the methodology and models implemented in this study can potentially be generalised to related domains, such as monitoring of medical devices, and geopolitical zones with region-specific product defect terminologies.

Although the classification terminology used in this study is highly specific to the regulatory environment in Singapore, there is potential in adapting or testing the

model in different regulatory environments or healthcare systems. For instance, a mapping of the classification terminology between MedDRA-HSA and the ontology of interest used by the target regulator could be discussed, which would allow the model to be applied in the target setting. Nonetheless, there should still be careful consideration into how comparable the different terminologies are in terms of semantics and granularity, as the model might over-specify or over-simplify predictions, leading to artificial precision that might not be meaningful to the target setting. Within the local context in Singapore, the developed model could also potentially be used to enhance existing health monitoring systems. This additional technology would likely complement, rather than replace, current detection and reporting workflows.

Conclusion

The technology described in this study will be useful for drug regulatory agencies like HSA as it serves as a first-pass screening tool, standardising case triage and augmenting routing workflows for more precise case prioritisation. Standardised categorisation could also improve signal detection algorithms and trend analysis capabilities. These insights could be potentially valuable in identifying emerging safety and/or product quality patterns across the health product ecosystem (information that could also impact clinical decisions made by healthcare professionals). The methodology and models implemented in this study can potentially be generalised to related domains, such as monitoring of medical devices, and geopolitical zones with region-specific product defect terminologies.

Methods

Fig. 4 illustrates the end-to-end development pipeline described in this study.

MedDRA-HSA dictionary

The MedDRA-HSA dictionary is an adapted version of ICH's MedDRA ontology (v24.0; available since 1 March 2021) for localised use within HSA. Specifically, this dictionary was generated in two major steps. Firstly, all LLTs under the High Level Group Term (HLGT) of "Product quality, supply, distribution, manufacturing and quality system issues" (n=384) were systematically reviewed by the pharmacovigilance team in HSA for relevance. Additionally, terms related to "advertisement non-compliance" (n=2) and "lack of efficacy" issues (n=5) were included for review, although strictly speaking these are not typically considered defects of product quality attributes. The reason for including these additional terms was to allow the model to have the capability to detect such issues – especially during environmental scans – as these issues are still relevant as part of the holistic assessment of a product's overall quality.

All reviewed terms (n=391) were then grouped into meaningful categories based on semantic description of the issues and local relevance. This gave a total of 38 distinct terms, denoted as MedDRA-HSA LLTs. To systematically reduce the number of LLT labels for classification, the HSA pharmacovigilance team further grouped some of the 38 terms into broader categories. Specifically, the 4 LLTs related to advertising compliance issues were grouped into a single term "advertisement non-compliance"; product label (whether inner or outer label), leaflet, and batch number issues were reduced from 8 individual LLTs into 3 separate terms, namely "Product label issue impacting strength, dose and/or safety", "Product leaflet issue impacting strength, dose and/or safety", and "Product label/leaflet issue NOT impacting strength, dose and/or safety". This reduction led to a final number of 30 MedDRA-HSA LLTs. Supplementary Table S1 shows the 30 MedDRA-HSA LLTs and the ICH terms that are related to each LLT.

Dataset collection and labelling

To train and develop a deep learning model that classifies reports of health product quality defects, a corpus of 13,948 product defect reports collected between January 2010 and December 2021 were used. The types of products include medicines, vaccines, complementary health products, health supplements and cosmetics. Records were collected from multiple sources, including but not limited to, companies' product defect reports (n=875), manual or automated environmental scanning of overseas product defect alerts posted by international drug regulatory agencies (n=8,638), good manufacturing practice inspections, product quality surveillance programmes, local adverse event reports submitted to HSA, and enforcement activities conducted by HSA (n=110), information sharing via international regulatory working groups, such as the Pharmaceutical Inspection Convention and Pharmaceutical Inspection Co-operation Scheme (PIC/S), and the Association of Southeast Asian Nations Post Marketing Alert System (ASEAN PMAS) (n=3,822), local government agencies, public sources of consumer data, and

patients' reports (n=237). Each report consists of a short title, typically between 5 and 20 words, and a description of the case of variable length. The title and the description contain the necessary information to classify the nature of the product defect reported and were used to train our system as described in the next sections. The corpus includes reports in non-English languages, mainly simplified and traditional Chinese and Malay. Reports in non-English languages were translated to English as described in the next section.

Training a machine learning system for classification requires providing samples of input data and the corresponding real class (ground truth) to the model. During training the model learns abstract features from the provided examples in the form of numerical vectors that enable it to discriminate between the different classes. Subsequently, during inference, the model generalizes the learned relationships between features and classes to predict the most probable class label of previously unseen data. To develop the deep learning classification system described in this study, each of the 13,948 collected reports were labelled with one of the 30 MedDRA-HSA LLT terms of product defects (Supplementary Table S1). To ensure consistency in the reference labels (ground truth), all the reports were revised and manually annotated for this study by a panel of domain experts following standardised criteria. These annotators were pharmacists trained and experienced in the areas of pharmacovigilance and the management of substandard medicine issues. They were actively handling such cases for at least 3 years at the time of annotation, signifying their currency in the subject domain. Out of the original 30 MedDRA-HSA categories, 9 were excluded from this study due to the insufficient availability of samples (less than 50 per class). The 9 categories excluded were: "Contamination with body fluid", "Product secondary packaging issue", "Product adhesion issue", "Product volume incorrect", "Product quantity incorrect", "Product storage issue", "Inappropriate release of product for distribution and distribution issue", "Distribution non-compliance (others)", and "Distribution non-compliance (non-registrable product)". The remaining 21 classes were used to develop our deep learning classifiers. Supplementary Table S8 shows examples of specific product defects for the 21 categories of MedDRA-HSA terms. The curated dataset comprised a total of 13,830 product defect reports accounting for 99.2% of all the collected reports. Table 2 shows the distribution of the 13,830 reports across the 21 class labels. The dataset is significantly imbalanced, with the majority class (Product adulterated and/or contains prohibited substance) representing 29.4% of the cases, and the minority class (Product formulation issue) representing only 0.58% of the cases.

Besides the case title, description, and class label, our dataset also contains other metadata of the reports, including a unique case identifier, the dates when the case was first reported and received by HSA, the level of severity, and the source of the report. Both the reporting source and date were used for data pre-processing and partitioning, respectively. On one side, reports received from three sources ("Company – Product Defect Report", "International work group – PIC/S", "International work group –ASEAN PMAS") follow a standard reporting template whereas the rest of reports received from other sources do not follow any specific format. The reports with a standard template underwent an additional pre-

processing step as described in the next subsection. On the other side, dates were used for a stratified partition of the data as described in the next sections.

To validate our model with prospective data, product defect reports received by HSA between January and June 2022 (n = 463) were manually labelled following the same criteria used to annotate the retrospective dataset. Prospective data comprised reports reported routinely to HSA and collected similarly to the reports used for model development. It was reasonably fair to assume that prospective reports (January to June 2022) were comparable in terms of quality and relevance to the retrospective reports used in this study (January 2010 to December 2021). However, due to the relatively small size of the prospective dataset, 5 categories of MedDRA-HSA terms were not represented.

Data pre-processing pipeline

The text of the reports was pre-processed to prepare the input data for model training. The following processing steps were applied:

1. Case descriptions in non-English languages were translated to English using the python libraries pylangtools for traditional Chinese and googletrans for other languages.
2. Each case description was cleaned by identifying and removing sentences that contain irrelevant information for the classification task, such as url sites, email addresses, and commonly reused expressions. Punctuation, casing, and digits were kept as in the original documents.
3. Reports from the three sources that follow standard reporting templates (see previous section) were further pre-processed by extracting relevant text from specific sections of the reporting template and discarding non-relevant and repetitive sections. These reports account for 5.6% of all the cases.
4. Finally, case title and case description were combined into a single attribute by inserting a period punctuation mark as separator.
5. To prepare the pre-processed records in a suitable format to feed into the model, the text from the previous step was tokenized, i.e., split into its basic building blocks, and the special token [CLS] was inserted at the beginning of each document. To ensure that all tokenized documents had the same length (512 tokens), short documents were padded with a special padding token whereas long documents were truncated. For further details on the tokenizer and [CLS] token, refer to the section Models and tokenizers.

Data partition

Machine learning models are first developed and subsequently evaluated with non-overlapping datasets. The development set is usually partitioned into a training set used to teach the model to recognize discriminatory patters in the data and a validation set used to optimize model hyperparameters by iterative cycles of training and validation. The test set is eventually used to evaluate the optimized model after development with previously unseen data. For the process to be successful data must be split evenly to ensure that training, validation, and test sets have the same or similar distribution of class labels. One strategy to achieve this

outcome is stratified sampling. In this strategy, the original dataset is segmented by class label. Subsequently, each segment is randomly sampled proportionally to their size in the dataset to preserve the same ratio of class labels in the training, validation, and test sets.

To minimise the impact of the COVID-19 pandemic on potential data drifts, the dataset was first split into pre-COVID (2010 to 2019 cases) and post-COVID (2020 to 2021 cases) groups. Due to the severe class imbalance, each subset was then partitioned into development and test sets in an 8:2 ratio using a stratified strategy to preserve the relative proportion of class labels in both datasets. The development set was further partitioned in an 8:2 ratio into the training and validation sets by stratified sampling. The resulting partitions from the pre-COVID and post-COVID groups were subsequently combined to obtain the final development (11,063 reports), training (8,849 reports), validation (2,214 reports), and test (2,767 reports) sets (Table 2). The training and validation sets were used to tune the models hyperparameters, whereas the development and test sets were used to train and evaluate the model with the optimal set of hyperparameters.

Models and tokenizers

Our product defects report classification system consists of two modules: a feature extractor and a classifier. The function of the feature extractor is to convert the text of the report into a dense numerical vector of 768 dimension that represents the contents of the document. The function of the classifier is to map the feature vector of the document to a particular MedDRA-HSA term. To extract features from text the BERT base architecture was selected due to 2 main reasons: 1) Relatively small scale (110 Million parameters, 12 encoder layers, and 12 attention heads) as compared to other language models with billions of parameters and hundreds of layers; 2) demonstrated SOTA performance in NLP classification tasks[27,29,32,33].

The original BERT architecture was designed to perform masked language modelling and next sentence prediction tasks[23]. To adapt this architecture to our text classification task, we replaced the language modelling head by a classification head. The classifier consisted of a shallow neural network containing one hidden layer of size 768 with linear activation and one output layer of size 21 (the number of MedDRA-HSA class labels) with softmax activation. Both layers were connected by an additional dropout regularisation layer.

To prepare the pre-processed text for model ingestion, the pre-trained BERT tokenizer was used. BERT tokenizer apply WordPiece, a tokenization algorithm that segments strings of text into subwords based on their relative frequencies[64]. The vocabulary of BERT comprises 30,522 tokens, including special tokens. BERT tokenizer always inserts the special [CLS] token at the start of each document. The function of this token if to capture a numerical vector representation of the entire document that can be subsequently used for text classification[23]. The maximum input length of BERT is 512 tokens. The BERT variant used in this study is case-sensitive.

Model tuning, training, and evaluation

Training LLMs from scratch is computationally expensive and requires large amounts of data. Instead, the most common paradigm employed to adapt LLMs to specific tasks is fine-tuning. Fine-tuning consists of taking a model pre-trained for general language tasks and re-training it with domain and task-specific data. During fine-tuning, the parameters of the algorithm are updated to minimize the error between the model predictions and the true labels. Pre-trained BERT[23] was fine-tuned using our curated dataset of health product defect reports. Besides the computational time and power savings, fine-tuning foundation models such as BERT has the advantage of an optimised initialisation of the model parameters and a quick convergence[61,65,66].

Pre-trained BERT base (case-sensitive) was fine-tuned using the training and validations datasets. Multiple rounds of fine-tuning were conducted to optimise the following hyperparameters: learning rate, number of training steps, and batch size. The training set was used to update the model parameters to maximize its performance on the product defect report classification task, whereas the validation set was used to evaluate the fine-tuned model after each round and adjust the hyperparameters. Model parameters were optimised using the Adam algorithm with weight decay, an implementation of the adaptive gradient algorithm Adam[67] with weight decay regularisation[68]. The models were tuned to minimise the cross-entropy loss between the standard reference labels and the predicted labels. To reduce overfitting, the optimiser weight decay rate and the kernel regularisation rate of the classification head were also tuned. After tuning, models were re-trained with the entire development set using the optimised hyperparameter values, and subsequently evaluated using the test set.

Baseline

To benchmark the performance of our model against baselines, we implemented and tested two probabilistic models. The first baseline assigns a MedDRA-HSA label to a test report based on the relative frequencies of each label in the training dataset. The second baseline was built by extracting the frequency counts of the words in each MedDRA-HSA label in the training dataset and subsequently training a Naïve Bayes classifier. This algorithm assigns a MedDRA-HSA label to a test report based on the probability of the observed word frequencies in the test report.

In addition, to contextualise performance against alternative BERT models specialised in the Scientific and Clinical domains, we fine-tuned SciBERT, Bio_ClinicalBERT and Bio_Discharge_Summary_BERT leveraging the same dataset and methodology used to fine-tune MedDefects-BERT:

- SciBERT: Initialised from BERT base and trained on a corpus of papers taken from Semantic Scholar[69].
- Bio_ClinicalBERT: Initialised from BioBERT and trained on all notes from MIMIC III dataset[70,71]. BioBERT was in turn initialised from BERT base and trained on large-scale biomedical corpora[72].
- Bio_Discharge_Summary_BERT: Initialised from BioBERT and trained only on discharge summaries from MIMIC III dataset[70].

Performance evaluation

Overall performance

The performance of our models was reported using standard metrics for multiclass classification tasks, including Precision, Recall, and F1 score for each class. Since these metrics are sensitive to class imbalance, we also reported their weighted and macro average scores. In addition, we reported other global performance scores such as Accuracy.

Positive and negative performance

To evaluate the effect of 1) the class label and 2) the report source on the positive and negative performance of MedDefects-BERT during batch inference, we conducted two spiking experiments. Spiking experiments refer to a method used to test the performance and scalability of a system under a sudden and significant increase in workload. This is often done by simulating a large, temporary surge in data volume. The two spiking experiments were conducted as follows:

1) Spiking experiment to assess positive and negative performance on the 21 MedDRA-HSA terms

This experiment assesses whether the MedDRA-HSA terms of the reports in a batch affect model performance irrespective of the sources. The following steps were performed:

- First, from each MedDRA-HSA term, a subset of the test dataset comprising 50% of the reports was randomly sampled (pre-spiked set).
- Subsequently, the pre-spiked set was injected with equal number of positive (same MedDRA-HSA term) or negative (any of the remaining 20 terms) reports randomly sampled from the test dataset to generate the post-spiked positive set and the post-spiked negative set, respectively.
- To obtain the baseline performance of the model on the pre-spiked set, batch inference was performed on this subset and the pre-spiking performance was evaluated by computing the Accuracy score on each MedDRA-HSA term (column “pre-spiking” in Supplementary Table S4).
- Finally, to obtain the post-spiking performance, batch inference was performed on the post-spiked positive and post-spiked negative sets and the Accuracy score was computed on each MedDRA-HSA term of both subsets (columns “post-spiking positive class” and “post-spiking negative class” in Supplementary Table S4).

2) Spiking experiment to assess positive and negative performance on the 5 sources of product defects reports

This experiment assesses whether the sources of the reports in a batch affect model performance irrespective of the MedDRA-HSA terms. The same methodology described in experiment 1) was followed, except for the reports being grouped and evaluated by source instead of MedDRA-HSA term (Supplementary Table S5).

Top-k multi-label performance

To simulate multi-label prediction scenarios, where minimisation of type II errors is prioritised at the cost of increasing ambiguity, we reported performance metrics for the top-1, top-2, and top-3 most probable predictions of the model. In addition, to simulate high-recall prediction scenarios, where recall of high impact cases is prioritised at the cost of increasing uncertainty, we reported performance metrics at different confidence threshold values for the subset of predictions with confidence above the threshold. We also reported the increments in F1 score and Recall between different models and prediction scenarios.

Error analysis

Out of the 397 reports misclassified by MedDefects-BERT (14% of the test set), 231 (58%) were randomly sampled to perform error analysis. Report title and description were manually inspected to identify possible sources of misclassification errors and categorised into four distinct groups by comparing the model output with the ground truth label ("Presence of specific keywords influencing predicted label", Overlap in label semantics", "Multiple related defects", and "Insufficient information"). Reports without any obvious reason for misclassification were assigned to an "Unclear reasons for misclassification" category (Table 3).

Correlation analyses

The correlation of top-1, top-2, and top-3 F1 scores for individual MedDRA-HSA categories with sample size in the development set was analysed by linear regression. The sample size (n) was 21 (the number of class labels). Pearson correlation coefficient (r), 95% confidence intervals, region of acceptance of the null hypothesis, and p-value were reported.

The correlation of median confidence scores for the top-1 predictions in individual MedDRA-HSA categories with F1 score was analysed by linear regression. The sample size (n) was 21 (the number of class labels). Pearson correlation coefficient, (r), 95% confidence intervals, region of acceptance of the null hypothesis, and p-value were reported.

Statistical tests

The significance in the observed differences in performance metrics under various scenarios was assessed by non-parametric one-tailed Wilcoxon signed-rank test and sign test. Non-parametric tests were selected because the distribution of the population data is not assumed to be normal. The Wilcoxon signed-rank test assesses the difference in the median between 2 populations, whereas the sign test assesses the differences in sign. Both the Wilcoxon signed-rank test and the sign test use 2 matched pairs of observations, such as the 21 class labels or the 5 sources of reports in our study. Two types of tests were conducted:

1) When assessing differences in performance across the 21 MedDRA-HSA terms, one-tailed tests were conducted to test the alternative hypothesis that the median F1 score and Recall in the new models and prediction scenarios are higher than in the baseline model (Table 1, Supplementary Tables S2 and

S7) or that the median Accuracy score in the post-spiking sets is lower than in the pre-spiking set (Supplementary Table S4). When testing the first hypothesis, the sample size (n) was always 21, whereas when testing the second hypothesis the sample size (n) was 16 when injecting positive samples and 20 when injecting negative samples. The reduction in sample size in the spiking test occurs because the Accuracy score differences between the pre-spiking and post-spiking sets for some MedDRA-HSA terms are null and the test discards those pairs.

2) When assessing differences in performance across the 5 sources of product defect reports, one-tailed tests were conducted to test the alternative hypothesis that the median Accuracy score in the post-spiking sets is lower than in the pre-spiking set (Supplementary Table S5). In those tests the sample size (n) was 5.

To run the sign test, the Z statistic was always used to test the hypothesis. However, to run the Wilcoxon signed-rank test, two types of statistics were used to evaluate the hypothesis, depending on the sample size (n):

- When n was higher than 10, the Z statistic was used.
- When n was lower than 10, the Z statistic does not qualify to test the hypothesis and the W value was used instead.

In both tests the corresponding p-values or the critical value for W (at $p < 0.05$) were reported.

Interpretability analysis

A) Global feature analysis

To understand the usefulness of the numerical vectors extracted by our system to represent the different categories of product defects described in the reports, and therefore to predict their MedDRA-HSA class labels, we visualised and compared the representations of multiple documents belonging to different categories in a low dimensional vector space. To perform classification, our system uses the last hidden state (feature map) of the [CLS] token output by the feature extractor as a vector representation of the full document[23]. To visualise the distance between vectors belonging to different class labels, the [CLS] vectors from the 8 MedDRA-HSA product defect categories with the highest F1 scores were projected into a lower dimensional space (2D) using the UMAP algorithm[57] and colour coded according to their standard reference labels.

B) Local feature analysis

To identify the segments of text that the model pays attention to when classifying a report into one category versus the rest and quantify their contributions (positive and negative) to the final model prediction, we performed a token influence analysis for multiple documents of each of the 21 MedDRA-HSA term categories. Individual-level (per sample) and aggregated-level (per class) token importance was derived by computing their Shapley values[58,73]. Briefly, documents were iteratively masked at random positions and passed to the model for prediction. Per sample token

importance was scored by averaging the influence of the masked tokens on the model output (i.e. the difference between the output of the unmasked and the masked documents) throughout all iterations for each document individually. Per class token importance was computed as the average Shapley value of the tokens in all the documents within each category of product defects. Per sample token importance analysis was illustrated by overlaying a gradient map of the Shapley values to the corresponding tokens in a document sample and by a force plot displaying the relative contribution of the tokens to the model output. Per class token importance analysis was reported by retrieving the top-10 full-word tokens with the highest average Shapley value for each class label. Average values were computed from up to 400 documents in each category of the training dataset.

Deep prompt tuning

Although fine-tuning is the standard paradigm to adapt language models to specific tasks, recently there has been research trend to develop alternative learning strategies to reduce the computational cost of updating all the model parameters. One such strategy is prompt tuning or its more advanced variant deep prompt tuning. Prompt tuning consists in injecting a small fraction of trainable parameters to the model while keeping the original parameters untrainable. Unlike in fine-tuning, during prompt tuning only the injected parameters are updated (learnt), while the rest the model parameters are frozen. This strategy has been shown to achieve comparable levels of performance as fine-tuning in specific tasks, while saving significant computational resources.

Our fine-tuned model was further improved by prompt-based learning. We sequentially combined fine-tuning with deep prompt tuning in an attempt to synergise the outcomes. Deep prompt tuning was performed by prepending a short sequence of 20 soft tokens to each layer of our fine-tuned model and re-training the model using the development set for an additional 5 epochs (when loss function converged) to learn the optimal prompt embeddings while freezing the parameters of the feature extractor module[52]. The resulting model was evaluated using the test set as described in previous sections.

Resources

The models described in this study were implemented using the TensorFlow framework (version 2.4.0) (https://www.tensorflow.org/). The pre-trained model parameters were downloaded from Hugging Face (https://huggingface.co/). The source code for deep prompt tuning is available at https://github.com/THUDM/P-tuning-v2. Shapley scores were calculated using the SHAP library (version 0.41.0) (https://shap.readthedocs.io/en/latest/api.html). UMAPS were obtained using the UMAP library (version 0.5.3) (https://umap-learn.readthedocs.io/en/latest/index.html). Metrics for model evaluation were computed using the scikit-learn library (version 1.0.2) (https://scikit-learn.org/stable/). The MedDRA-HSA dictionary used in this study is available in Supplementary Table S1. The sign test and the Wilcoxon signed-rank test were run using the open access social science statistics calculators (https://www.socscistatistics.com/tests/signtest/default.aspx;

https://www.socscistatistics.com/tests/signedranks/default2.aspx). The Pearson correlation test was run using the open access statistics kingdom calculator (https://www.statskingdom.com/correlation-calculator.html)

Code availability

Open source libraries used in this study are referenced in the Resources section of the Methods. Custom code developed in this study is available at the following GitHub repository: https://github.com/hytting/Product-defect

References

1. Nagaich, U. & Sadhna, D. Drug recall: An incubus for pharmaceutical companies and most serious drug recall of history. *Int. J. Pharm. Investig.* **5**, 13-19 (2015).

2. US Food & Drug Administration. Annual Report (2022). https://www.fda.gov/media/166289/download.

3. Lindström-Gommers, L. & Mullin, T. International Conference on Harmonization: Recent Reforms as a Driver of Global Regulatory Harmonization and Innovation in Medical Products. *Clin. Pharmacol. Ther*. **105**, 926-931 (2019).

4. Ang, P. S. *et al.* A risk classification model for prioritising the management of quality issues relating to substandard medicines in Singapore. *Pharmacoepidemiol. Drug Saf.* **31**, 729–738 (2022).

5. Vasey, B. *et al.* Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat.Med.* **28**, 924–933 (2022).

6. Vaswani, A. *et al.* Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017-December**, 5999–6009 (2017).

7. Vig, J. A Multiscale Visualization of Attention in the Transformer Model. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of System Demonstrations* 37–42. arXiv preprint arXiv:1906.05714 (2019).

8. Rogers, A., Kovaleva, O. & Rumshisky, A. A Primer in BERTology: What we know about how BERT works. *Trans. Assoc. Comput. Linguist.* **8**, 842–866 (2020).

9. Clark, K., Khandelwal, U., Levy, O. & Manning, C. D. What Does BERT Look At? An Analysis of BERT's Attention. arXiv preprint arXiv:1906.04341 (2019).

10. He, P., Liu, X., Gao, J. & Chen, W. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. arXiv preprint arXiv:2006.03654 (2020).

11. Suzgun, M. *et al.* Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. arXiv preprint arXiv:2210.09261 (2022).

12. Raffel, C. *et al.* Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **21**, 5485-5551 (2020).

13. Radford, A. *et al.* Language Models are Unsupervised Multitask Learners. *OpenAI blog* **1**, 9 (2019).

14. Yang, Z. *et al.* XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Adv. Neural Inf. Process. Syst.* **32** (2019).

15. Liu, Y. *et al.* RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907*.11692* (2019).

16. Brown, T. B. *et al.* Language Models are Few-Shot Learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877-1901 (2020).

17. Clark, K. *et al.* ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. arXiv preprint arXiv:2003.10555 (2020).

18. Bahdanau, D., Cho, K. H. & Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. arXiv preprint arXiv:1409.0473 (2014).

19. Bommasani, R. et al. On the Opportunities and Risks of Foundation Models. arXiv preprint arXiv:2108.07258 (2021).

20. Ghaseminejad Raeini, M. The evolution of language models: From N-Grams to LLMs, and beyond. *Natural Language Processing Journal* **12**, 100168 (2025).

21. Hu, Y. *et al.* PheCatcher: Leveraging LLM-Generated Synthetic Data for Automated Phenotype Definition Extraction from Biomedical Literature. *Stud. Health Technol. Inform.* **329**, 718–722 (2025).

22. Li, Y., Li, J., He, J. & Tao, C. AE-GPT: Using Large Language Models to extract adverse events from surveillance reports-A use case with influenza vaccine adverse events. *PLoS One* **19**, (2024).

23. Devlin, J. *et al.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *N4ACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* **1**, 4171–4186. arXiv preprint arXiv:1810.04805 (2019).

24. Sun, C. *et al.* Biomedical named entity recognition using BERT in the machine reading comprehension framework. *J Biomed. Inform.* **118**, (2021).

25. Gu, Y. U. *et al.* Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthc. (HEALTH)* **3**.1, 1-23 (2021).

26. Tan, F. *et al.* MGEL: Multigrained Representation Analysis and Ensemble Learning for Text Moderation. *IEEE Trans. Neural Netw. Learn. Syst.* **34**, 7014-7023 (2022).

27. Senn, S., Tlachac, M. L., Flores, R. & Rundensteiner, E. Ensembles of BERT for Depression Classification. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* **2022**, 4691–4694 (2022).

28. Widad, A., El Habib, B. L. & Ayoub, E. F. Bert for Question Answering applied on Covid-19. *Procedia. Comput. Sci.* **198**, 379–384 (2022).

29. Xu, C., Yuan, F. & Chen, S. BJBN: BERT-JOIN-BiLSTM Networks for Medical Auxiliary Diagnostic. *J. Healthc. Eng.* **2022**, (2022).

30. Ji, Z., Wei, Q. & Xu, H. BERT-based Ranking for Biomedical Entity Normalization. *AMIA Jt. Summits Transl. Sci. Proc.* **2020**, 269–277 (2020).

31. Jiang, L. *et al.* IUP-BERT: Identification of Umami Peptides Based on BERT Features. *Foods* **11**, 3742 (2022).

32. Aldahdooh, J., Vähä-Koskela, M., Tang, J. & Tanoli, Z. Using BERT to identify drug-target interactions from whole PubMed. *BMC Bioinformatics* **23**, 245 (2022).

33. Tejani, A. S. *et al.* Performance of Multiple Pretrained BERT Models to Automate and Accelerate Data Annotation for Large Datasets. *Radiol. Artif. Intell.* **4**, (2022).

34. Kuo, C. C., Chen, K. Y. & Luo, S. B. Audio-Aware Spoken Multiple-Choice Question Answering with Pre-Trained Language Models. *IEEE/ACM Trans. Audio Speech Lang. Process.* **29**, 3170–3179 (2021).

35. Wang, Z. Y. *et al.* Pre-Trained Models Based Receiver Design With Natural Redundancy for Chinese Characters. *IEEE Communications Letters* **26**, 2350–2354 (2022).

36. Kowsher, M. *et al.* Bangla-BERT: Transformer-Based Efficient Model for Transfer Learning and Language Understanding. *IEEE Access* **10**, 91855–91870 (2022).

37. Zhu, X., Wu, H. & Zhang, L. Automatic Short-Answer Grading via BERT-Based Deep Neural Networks. *IEEE Transactions on Learning Technologies* **15**, 364–375 (2022).

38. Liu, N., Hu, Q., Xu, H., Xu, X. & Chen, M. Med-BERT: A Pretraining Framework for Medical Records Named Entity Recognition. *IEEE Trans. Industr. Inform.* **18**, 5600–5608 (2022).

39. Zhou, C. Comparative Evaluation of GPT, BERT, and XLNet: Insights into Their Performance and Applicability in NLP Tasks. *Transactions on Computer Science and Intelligent Systems Research* **7**, 415–421 (2024).

40. Gardazi, N. M. *et al.* BERT applications in natural language processing: a review. *Artificial Intelligence Review 2025* **58**, 166 (2025).

41. Zhong, R., Ghosh, D., Klein, D. & Steinhardt, J. Are Larger Pretrained Language Models Uniformly Better? Comparing Performance at the Instance Level. *Findings of the Association for Computational Linguistics: ACL-IJCNLP* **2021**, 3813–3827 (2021).

42. Vinyals, O.*et al.* Matching Networks for One Shot Learning. *Adv. Neural Inf. Process Syst.* **29,** 3637–3645 (2016).

43. Baevski, A. *et al.* Cloze-driven Pretraining of Self-attention Networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference* 5360–5369. arXiv preprint arXiv:1903.07785 (2019).

44. Schick, T. & Schütze, H. Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference. *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference* 255–269. arXiv preprint arXiv:2001.07676 (2020).

45. Schick, T. & Schütze, H. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference* 2339–2352. arXiv preprint arXiv:2009.07118 (2020).

46. Gao, T., Fisch, A. & Chen, D. Making Pre-trained Language Models Better Few-shot Learners. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference* 3816–3830. arXiv preprint arXiv:2012.15723 (2020).

47. Shin, T., Razeghi, Y., Logan, R. L., Wallace, E. & Singh, S. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* 4222–4235. arXiv preprint arXiv:2010.15980 (2020).

48. Lester, B., Al-Rfou, R. & Constant, N. The Power of Scale for Parameter-Efficient Prompt Tuning. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings* 3045–3059. arXiv preprint arXiv:2104.08691 (2021).

49. Liu, X. *et al.* GPT Understands, Too. AI Open (2023). doi: https://doi.org/10.1016/j.aiopen.2023.08.012.

50. Li, X. L. & Liang, P. Prefix-Tuning: Optimizing Continuous Prompts for Generation. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference* 4582–4597. arXiv preprint arXiv:2101.00190 (2021).

51. Qin, G. & Eisner, J. Learning How to Ask: Querying LMs with Mixtures of Soft Prompts. *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference* 5203–5212. arXiv preprint arXiv:2104.06599 (2021).

52. Liu, X. *et al.* P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks. arXiv preprint arXiv:2110.07602 (2021).

53. Khandelwal, U., He, H., Qi, P. & Jurafsky, D. Sharp Nearby, Fuzzy Far Away: How Neural Language Models Use Context. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)* **1**, 284–294. arXiv preprint arXiv:1805.04623 (2018).

54. Zorzi, M., Combi, C., Lora, R., Pagliarini, M. & Moretti, U. Automagically encoding Adverse Drug Reactions in MedDRA. *2015 International Conference on Healthcare Informatics, IEEE* 90–99 (2015).

55. Tiftikci, M., Özgür, A., He, Y. & Hur, J. Machine learning-based identification and rule-based normalization of adverse drug reactions in drug labels. *BMC Bioinformatics* **20**, 1–9 (2019).

56. Létinier, L. *et al.* Artificial Intelligence for Unstructured Healthcare Data: Application to Coding of Patient Reporting of Adverse Drug Reactions. *Clin. Pharmacol. Ther.* **110**, 392-400 (2021).

57. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv preprint arXiv:1802.03426 (2018).

58. Lundberg, S. M. & Lee, S. I. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* **30** (2017).

59. Peryea, T. *et al.* Global Substance Registration System: consistent scientific descriptions for substances related to health. *Nucleic Acids Res.* **49**, D1179-D1185 (2021).

60. Li, Y. *et al.* Artificial intelligence-powered pharmacovigilance: A review of machine and deep learning in clinical text-based adverse drug event detection for benchmark datasets. *J. Biomed. Inform.* **152**, 104621 (2024).

61. Howard, J. & Ruder, S. Universal Language Model Fine-tuning for Text Classification. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)* **1**, 328–339. arXiv preprint arXiv:1801.06146 (2018).

62. He, J. *et al.* Prompt Tuning in Biomedical Relation Extraction. *J. Healthc. Inform. Res.* **8**, 206–224 (2024).

63. Chooi, W. H. *et al.* Vaccine contamination: Causes and control. *Vaccine* **40**, 1699–1701 (2022).

64. Wu, Y. *et al.* Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv preprint arXiv:1609.08144 (2016).

65. Hao, Y., Dong, L., Wei, F. & Xu, K. Visualizing and Understanding the Effectiveness of BERT. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference* 4143–4152. arXiv preprint arXiv:1908.05620 (2019).

66. Tan, C. *et al.* A Survey on Deep Transfer Learning. In: Kůrková, V., Manolopoulos, Y., Hammer, B., Iliadis, L., Maglogiannis, I. (eds) Artificial Neural Networks and Machine Learning – ICANN 2018. ICANN 2018. Lecture Notes in Computer Science(), vol 11141. Springer, Cham. https://doi.org/10.1007/978-3-030-01424-7_27.

67. Kingma, D. P. & Ba, J. L. Adam: A Method for Stochastic Optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. arXiv preprint arXiv:1412.6980 (2014).

68. Loshchilov, I. & Hutter, F. Decoupled Weight Decay Regularization. *7th International Conference on Learning Representations, ICLR 2019*. arXiv preprint arXiv:1711.05101 (2017).

69. Beltagy, I., Lo, K. & Cohan, A. SciBERT: A Pretrained Language Model for Scientific Text. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference* 3615–3620 (2019).

70. Alsentzer, E. *et al.* Publicly Available Clinical BERT Embeddings. *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 72-78 (2019).

71. Johnson, A. E. W. *et al.* MIMIC-III, a freely accessible critical care database. *Scientific Data 2016* **3**, 1–9 (2016).

72. Lee, J. *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2019).

73. Shrikumar, A., Greenside, P. & Kundaje, A. Learning Important Features Through Propagating Activation Differences. *34th International Conference on Machine Learning ICML 2017* **70***,* 3145-3153 (2017).

Acknowledgements

We acknowledge Michelle Ng, Chih Tzer Choong, Doris Phuah, Dorothy Tan, Filina Tan, Huilin Huang, Maggie Tan and Jalene Poh for their expert opinion and assistance in this work.

We want to thank Govindaraj Roshni Daksha for performing a thorough error analysis and providing valuable insights.

Funding

This initiative received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author contributions

Desmond Teo, Sreemanee Dorajoo and Pei San Ang proposed the research idea. Vicente Sancenon, Yiting Huang and Lin Zou designed the models and analysed the data. Desmond Teo, Sreemanee Dorajoo and Pei San Ang provided the domain expertise for manual annotation of the data. Han Leong Goh and Andy Ta provided the thought leadership for the project.

Data availability

The data that support the findings of this study are not openly available due to the confidentiality and proprietary nature of the records, especially those obtained from companies' product defect reports, information sharing via international regulatory working groups, and good manufacturing practice inspections. Data are located in controlled access data storage at the Singapore Health Sciences Authority. The data are, however, available from the corresponding author upon reasonable request.

Additional Information

Competing interests

The authors declare no competing interests.

Figures and Tables

Figure 1. Performance and confidence analyses. **a.** Top-1 (blue), top-2 (orange), and top-3 (green) F1 scores for individual MedDRA-HSA categories as a function of sample size in the development set. A linear model was fit by regression to the data points. Solid lines represent the best fitted models for each top-k group. Shaded areas represent 95% confidence intervals. Top-1 (Pearson's r: 0.643; 95% CI: [0.2921, 0.8411]; p-value: 0.0016), top-2 (Pearson's r: 0.735; 95% CI: [0.4439, 0.8856]; p-value: 0.0001), top-3 (Pearson's r: 0.635; 95% CI: [0.2808, 0.8374]; p-value: 0.0020). **b.** Median confidence scores for the top-1 predictions in individual MedDRA-HSA categories as a function of F1 score. A linear model was fit by regression to the data points. Solid lines represent the best fitted model. Shaded areas represent 95% confidence intervals (Pearson's r: 0.927; 95% CI: [0.8253, 0.9703]; p-value <0.00001). **c**. Boxplot distribution of the confidence scores for the top-1 correct (orange) and incorrect (blue) predictions for individual MedDRA-HSA categories ranked by increasing F1 scores. The upper and lower edges of the boxes represent the 25^{th} and 75^{th} percentiles, respectively, and the central line represents the median. **d.** Model performance metrics as a function of increasing confidence thresholds. To illustrate the impact of the threshold on recall, this metric is shown for all predictions and for predictions above the confidence threshold. Blue: accuracy; orange: percent of predictions with confidence below threshold; green: precision; dashed red: recall (all predictions); solid red: recall (predictions with confidence above threshold). Vertical dotted line indicates threshold for optimal recall.

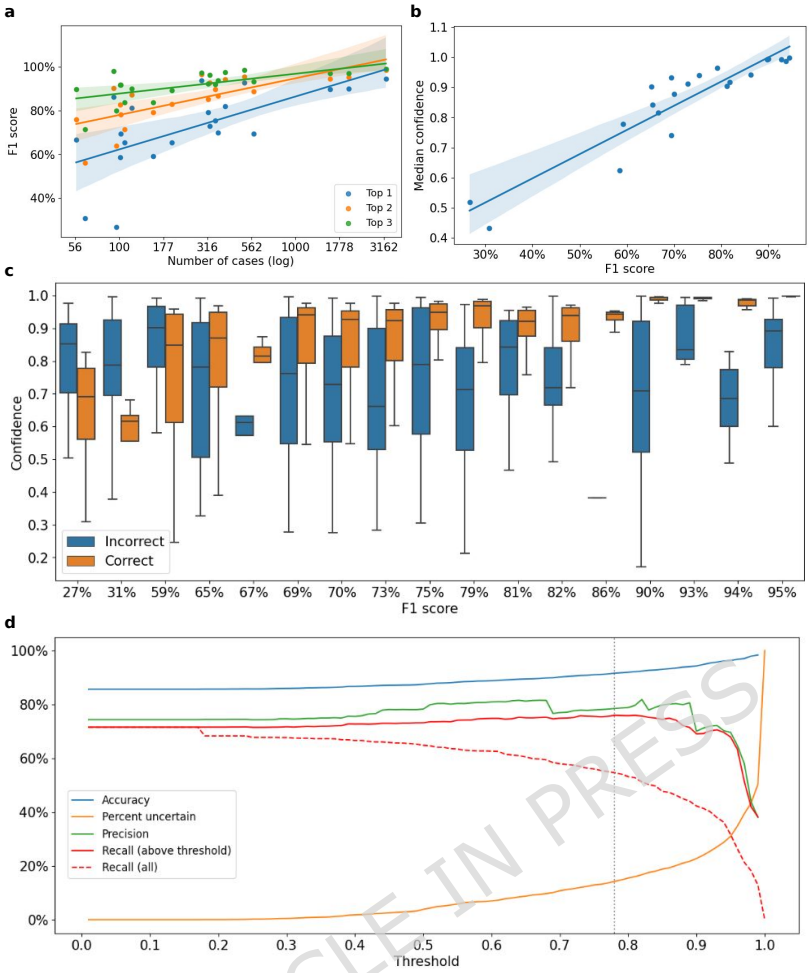

Figure 2. Global (document-level) feature analysis. UMAP projection of the feature vectors (document representations) from reports in the 8 most common MedDRA-HSA categories. Colours represent the standard reference label (ground truth) of the report as follows:

- Product adulterated and/or contains prohibited substance
- Out of specification or out of trend test result
- Manufacturing non-compliance
- Contamination with microbes
- Contamination with foreign matter
- Product counterfeit
- Product expiration date missing, illegible or incorrect
- Contamination with glass and/or metal particle
- Other categories (only for predictions)
- ○ Correctly classified reports
- ⊠ Misclassified reports

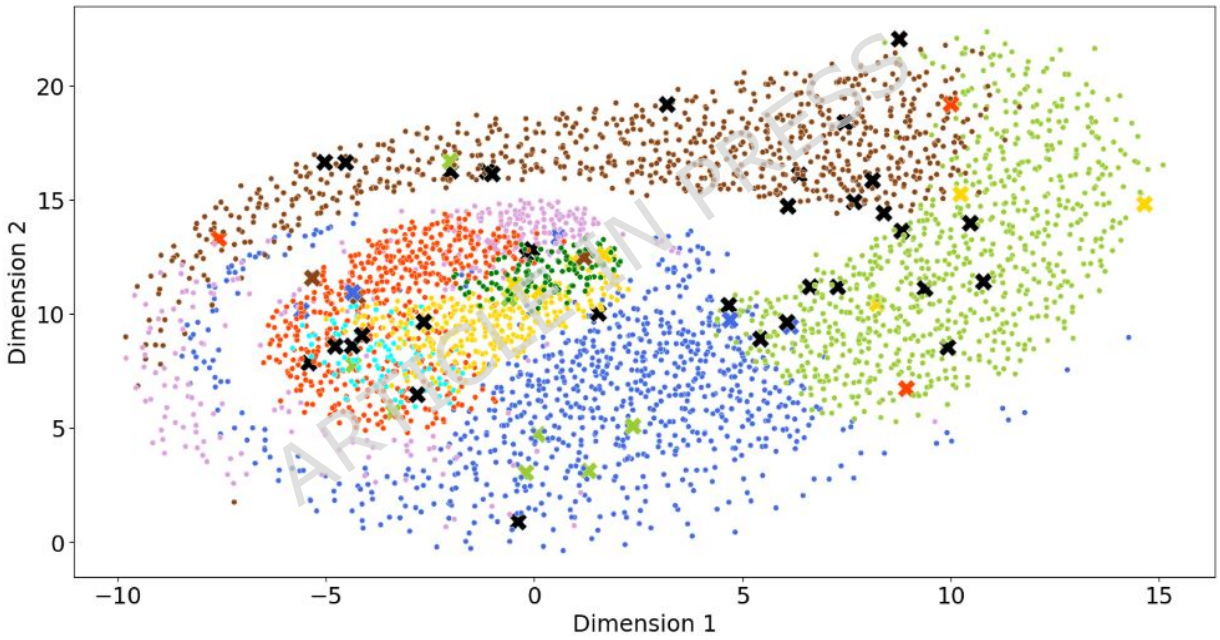

Figure 3. Local (token-level) feature analysis. Example of instance-level SHAP analysis for a single report correctly predicted in the category of “Contamination with microbes” with a confidence score of 0.69. **a.** Force plot displaying the relative influence of all the document tokens to the final predicted probability. The magnitude of each token importance is proportional to the size of its chevron arrow. **b.** Gradient map highlighting the relative contribution of each token in the document to the final prediction. The magnitude of the contribution is proportional to the intensity of the colour. Sensitive words referring to drugs or businesses have been replaced by generic names. Red: positive influence; Blue: negative influence.

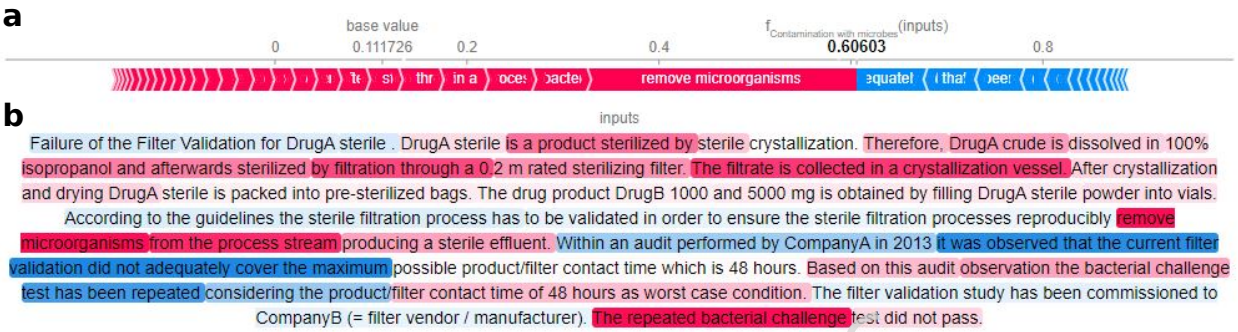

Figure 4. Schematic diagram of the model development and evaluation flow performed in this study. A total of 13,948 Health product defect reports were collected from January 2010 to December 2021 from multiple sources, including web sites, company reports and alerts raised by pharmaceutical inspection authorities (PIC/S). Records were coded by a panel of domain experts into 30 different categories following MedDRA-HSA terminology. Minority classes were removed from the study and the remaining 13,830 reports belonging to 21 categories were used to train our system by a two-step methodology. In the first step (fine-tuning) the weights of all the layers of a pre-trained BERT model plus a newly initialised classification head were updated during training. In the second step (deep prompt-tuning) learnable soft prompts were prepended to all the model layers and updated during training while freezing the rest of the model parameters. After development, performance and interpretability analyses were conducted to evaluate the system. See Methods section for further details. PIC/S: Pharmaceutical Inspection Convention and Pharmaceutical Inspection Co-operation Scheme.

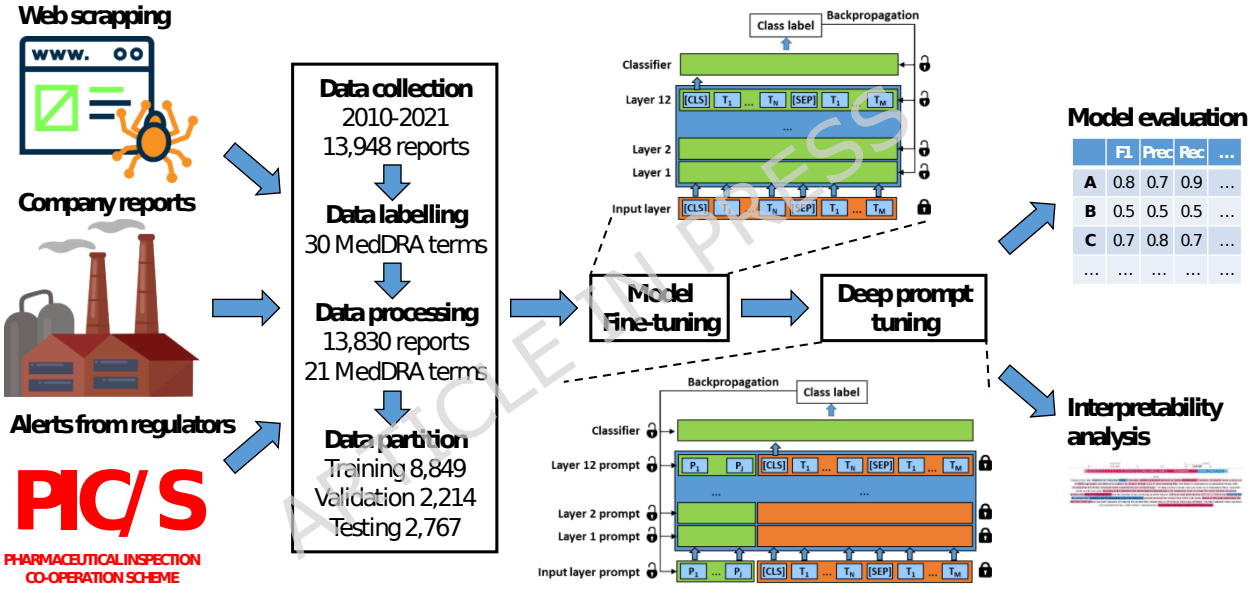

Table 1. Performance metrics of various models developed in this study. Comparison of the performance metrics of the various models developed and described in this study. Every model is identified by the following set of features: Fine-tuning/Prompt-tuning: the paradigm used to train the model; No threshold/Confidence threshold: Post-processing method used to generate the final model output; top-1; top-2; top-3: the number of most probable categories predicted by the model to compute the metrics (see Methods for further details). Precision, Recall, and F1 score were computed on the test set for each individual MedDRA term and aggregated as macro and weighted average scores for each severity group and across all the terms. In addition, Accuracy was reported for each severity group and overall. Increments (Δ) in F1 score and Recall between selected models were also computed as described in the main text. The significance in the observed increments were assessed by non-parametric one-tailed Wilcoxon signed-rank test and sign test, and the corresponding Z statistic and p-values were reported. F1: F1-score; Prec: Precision; Rec: Recall; Sup: Support; ***: p-value < 0.00001. ΔF1* and ΔRec*: increment in F1-score and Recall relative to top-1; ΔF1$^{\#}$ and ΔRec$^{\#}$: increment in F1-score and Recall relative to No threshold; ΔF1$^{\$}$ and ΔRec$^{\$}$: increment in F1-score and Recall relative to Fine-tuning.

			Fine-tuned (MedDefects-BERT)														Deep prompt tuned (MedDefects-DPT-BERT)				
			No threshold										Confidence threshold: 0.78					No threshold				
			Top-1			Top-2				Top-3					Top-1					Top-1				
Severity	MedDRA-HSA term	Sup	F1	Prec	Rec	F1	ΔF1*	Prec	Rec	ΔRec*	F1	ΔF1*	Prec	Rec	ΔRec*	F1	ΔF1#	Prec	Rec	ΔRec#	F1	ΔF1$	Prec	Rec	ΔRec$
High	Contamination with glass and/or metal particle	30	81%	76%	87%	87%	6%	84%	90%	3%	90%	9%	90%	90%	3%	86%	5%	81%	91%	4%	82%	1%	81%	83%	-4%
	Contamination with microbes	128	93%	95%	91%	96%	3%	98%	93%	2%	98%	5%	99%	98%	7%	94%	1%	97%	92%	1%	91%	-2%	91%	91%	0%
	Product counterfeit	73	94%	96%	92%	96%	2%	100%	93%	1%	97%	3%	100%	95%	3%	97%	3%	97%	97%	5%	93%	-1%	96%	90%	-2%
	Product formulation issue	16	31%	40%	25%	56%	25%	78%	44%	19%	71%	40%	83%	62%	37%	0%	-31%	0%	0%	-25%	23%	-8%	30%	19%	-6%
	Product label issue impacting strength, dose and/or safety	146	69%	65%	74%	89%	20%	85%	92%	18%	93%	24%	92%	95%	21%	80%	11%	78%	82%	8%	72%	3%	73%	70%	-4%
	Product mix up	50	65%	69%	62%	83%	18%	84%	82%	20%	89%	24%	87%	92%	30%	73%	8%	76%	71%	9%	66%	1%	64%	68%	6%
	Accuracy		**79%**			**89%**	**10%**				**94%**	**14%**				**84%**	**5%**				**78%**	**-1%**			
	Macro		**72%**	**74%**	**72%**	**85%**	**12%**	**88%**	**82%**	**11%**	**90%**	**18%**	**92%**	**89%**	**17%**	**72%**	**-1%**	**72%**	**72%**	**0%**	**71%**	**-1%**	**73%**	**70%**	**-2%**
	Weighted		**79%**	**79%**	**80%**	**90%**	**11%**	**91%**	**89%**	**10%**	**94%**	**15%**	**94%**	**94%**	**14%**	**84%**	**5%**	**84%**	**84%**	**4%**	**79%**	**0%**	**80%**	**78%**	**-1%**
Medium	Accompanying dose delivery device issue	25	59%	75%	48%	83%	24%	90%	76%	28%	92%	33%	96%	88%	40%	48%	-11%	83%	33%	-15%	64%	5%	74%	56%	8%
	Advertisement non-compliance	99	82%	84%	80%	94%	12%	91%	97%	17%	97%	15%	97%	98%	18%	89%	7%	88%	91%	11%	82%	0%	82%	83%	3%
	Contamination with chemical substance	81	73%	81%	67%	93%	20%	99%	88%	21%	96%	23%	100%	93%	26%	79%	6%	82%	76%	9%	76%	3%	78%	74%	7%
	Contamination with foreign matter	79	79%	88%	72%	85%	6%	91%	80%	8%	92%	13%	95%	90%	18%	89%	10%	92%	85%	13%	83%	4%	83%	82%	10%
	Lack of efficacy	27	65%	64%	67%	71%	6%	69%	74%	7%	84%	19%	82%	85%	18%	60%	-5%	56%	64%	-3%	68%	3%	62%	74%	7%
	Manufacturing non-compliance	394	90%	86%	94%	95%	5%	93%	96%	2%	97%	7%	96%	97%	3%	95%	5%	92%	97%	3%	91%	1%	89%	93%	-1%
	Out of specification or out of trend test result	501	90%	90%	90%	95%	5%	95%	96%	6%	97%	7%	97%	97%	7%	94%	4%	93%	96%	6%	90%	0%	89%	90%	0%
	Product adulterated and/or contains prohibited substance	814	95%	94%	95%	98%	3%	98%	99%	4%	99%	4%	99%	99%	4%	96%	1%	97%	96%	1%	94%	-1%	95%	94%	-1%
	Product deposit	14	67%	69%	64%	76%	9%	73%	79%	15%	90%	23%	87%	93%	29%	88%	21%	78%	100%	36%	71%	4%	65%	79%	15%
	Product expiration date missing, illegible or incorrect	23	86%	79%	96%	90%	4%	82%	100%	4%	98%	12%	96%	100%	4%	88%	2%	78%	100%	4%	81%	-5%	72%	91%	-5%
	Product leaflet issue impacting strength, dose and/or safety	26	69%	74%	65%	78%	9%	90%	69%	4%	92%	23%	100%	85%	20%	83%	14%	91%	77%	12%	67%	-2%	73%	62%	-3%
	Product primary packaging issue	87	75%	77%	74%	90%	15%	90%	90%	16%	92%	17%	92%	92%	18%	81%	6%	79%	83%	9%	76%	1%	77%	76%	2%
	Product unregistered	91	70%	81%	62%	87%	17%	97%	78%	16%	94%	24%	98%	90%	28%	79%	9%	84%	74%	12%	70%	0%	73%	67%	5%
	Accuracy		**88%**			**94%**	**7%**				**97%**	**9%**				**92%**	**5%**				**88%**	**1%**			
	Macro		**77%**	**80%**	**75%**	**87%**	**10%**	**89%**	**86%**	**11%**	**94%**	**17%**	**95%**	**93%**	**18%**	**82%**	**5%**	**84%**	**82%**	**8%**	**78%**	**1%**	**78%**	**79%**	**4%**
	Weighted		**88%**	**88%**	**88%**	**94%**	**6%**	**95%**	**94%**	**7%**	**97%**	**9%**	**97%**	**97%**	**9%**	**92%**	**4%**	**92%**	**92%**	**4%**	**88%**	**0%**	**88%**	**88%**	**1%**

Table 2. Characteristics of the dataset used in this study. Dataset contains product defect reports collected between January 2010 and December 2021 from various sources and annotated for this study by a panel of experts following standard criteria to ensure consistency in the labels (see Methods for further details). Data was partitioned into the development and test sets in a 8:2 ratio and the development set was further partitioned into training and validation sets for model hyperparameter tuning also in a 8:2 ratio. Both partitions were performed using a stratified sampling strategy to preserve the relative proportion of MedDRA-HSA terms and pre/post-pandemic representation in all datasets.

Severity	MedDRA-HSA term	All samples	Development	Training	Validation	Testing
High	Contamination with glass and/or metal particle	148 (1.1%)	118	94	24	30
	Contamination with microbes	641 (4.6%)	513	410	103	128
	Product counterfeit	365 (2.6%)	292	233	59	73
	Product formulation issue	80 (0.6%)	64	52	12	16
	Product label issue impacting strength, dose and/or safety	725 (5.2%)	579	463	116	146
	Product mix up	249 (1.8%)	199	159	40	50
Medium	Accompanying dose delivery device issue	126 (0.9%)	101	81	20	25
	Advertisement non-compliance	497 (3.6%)	398	318	80	99
	Contamination with chemical substance	406 (2.9%)	325	260	65	81
	Contamination with foreign matter	397 (2.9%)	318	254	64	79
	Lack of efficacy	134 (1.0%)	107	86	21	27
	Manufacturing non-compliance	1965 (14.2%)	1571	1257	314	394
	Out of specification or out of trend test result	2507 (18.1%)	2006	1605	401	501
	Product adulterated and/or contains prohibited substance	4068 (29.4%)	3254	2603	651	814
	Product deposit	71 (0.5%)	57	46	11	14
	Product expiration date missing, illegible or incorrect	116 (0.8%)	93	74	19	23
	Product leaflet issue impacting strength, dose and/or safety	128 (0.9%)	102	81	21	26
	Product primary packaging issue	437 (3.2%)	350	280	70	87
	Product unregistered	455 (3.3%)	364	291	73	91
Low	Product label/leaflet issues NOT impacting strength, dose and/or safety	120 (0.9%)	96	77	19	24
	Product physical issue	195 (1.4%)	156	125	31	39
Total		**13830 (100.0%)**	**11063**	**8849**	**2214**	**2767**

Table 3. Error analysis. Categories, frequency, and examples of product defect report misclassification errors.

Category	Frequency of Error (n)	Description of Error Category	Case Examples (Explanations)
Presence of specific keywords influencing predicted label	50.6% (117)	Model prediction is primarily driven by the presence of certain keywords (which are strongly associated with certain labels) rather than the overall context, leading to misclassification	• *"…recall of [product] due to due to packaging mixed-up (red plastic flip offs used instead of grey plastic flips off)…"* (Ground truth is "Product primary packaging issue" while predicted label was "Product mix up". Specific mention of *"mixed-up"* may have led to misclassified label.) • *"..recall of [product] due to a potential for dose delivery out of specification…"* (Ground truth is "Accompanying dose delivery device issue" while predicted label was "Out of specification or out of trend test result". Specific mention of *"out of specification"* possibly led to erroneous label prediction)
Overlap in label semantics	24.2% (56)	Some MedDRA labels have closely related definitions where distinctions depend on subtle contextual differences	• *"[Product] contains beta phenyl gamma aminobutyric acid…"* (Ground truth is "Contamination with chemical substance" while predicted label was "Product adulterated and/or contains prohibited substance". In this case, chemical substance contamination is more factual and appropriate as there is no evidence of illicit activity, however this contextual information was not available to the model in the initial report.) • *"..voluntary recall of batches…in response to complaints of presence of visual particulate after reconstitution of the product. ..most likely root cause…prolonged contact time of product solution with the filters…"* (Ground truth is "Product deposit" while predicted label was "Contamination with foreign matter". In this case, the "*visual particulates"* were a result of interaction between the product solution and filters which were not external substances (foreign) to the product, however the model may have not assigned sufficient importance to this information to predict the label.)
Multiple related defects	10.0% (23)	Description of the defect mentions two or more related or sequential issues, making the report suitable to receive two or more labels	• "..*reports of [product]…compatibility issue with trace elements…potential for formation of a precipitate…"* (Ground truth is "Product formulation issue" while predicted label was "Product deposit". In this case, both labels are appropriate.) • "..*[product] recalled due to manufacturing defect with the needle…that may prevent the device from working properly."* (Ground truth is "Accompanying dose delivery device issue" while predicted label was "Manufacturing non-compliance". In this case, both labels are appropriate.)
Insufficient information	8.7% (20)	Description of the defect lacks sufficient detail for accurate classification	• *"Recall of [product] due to detritus contamination…"* (Ground truth is "Contamination with foreign matter" while predicted label was "Contamination with microbes". The mention of *"detritus"* only is insufficient for an accurate classification.) • *"[Company] recalls [product] due to potential contamination."* (Ground truth is "Contamination with foreign matter" while predicted label was "Contamination with microbes". Specific type of contamination is not mentioned in the report.)
Unclear reasons for misclassification	6.5% (15)	Description of the defect aligns with the ground truth and there are no obvious reasons for misclassification	• *"..voluntary recall of [product]…due to sub-potency. Potential adverse events…may include reduced effectiveness."* (Ground truth is "Lack of efficacy" while predicted label was "Out of specification or out of trend test result". In this case, the description is clearly linked to lack of efficacy of the product.) • *"Cancellations of [product]…due to non-compliance…failed to comply with a condition of listing…"* (Ground truth is "Manufacturing non-compliance" while predicted label is "Lack of efficacy". In this case, the description clearly states non-compliance while there is no indication about the product's effectiveness nor efficacy.)

Table 4. Global summary of token-level feature analysis. Top-10 full-word tokens with the highest positive average Shapley values across the 21 MedDRA-HSA labels. Shapley values were computed from up to 400 reports per class label in the training set (see Methods for further details). Tokens are shown in their original case, as a case-sensitive BERT was used in this study. Sensitive tokens referring to countries and cities have been replaced by generic names. ots: off tool sample; rdi: recommended daily intake; API: active pharmaceutical ingredient.

Severity	MedDRA-HSA term	Top-10 influential words
High	Contamination with glass and/or metal particle	metallic, aluminium, wire, iron, Glass, resembling, steel, glass, metal, flat
	Contamination with microbes	fungi, transmission, variant, pores, integrity, pus, Counts, fungus, micro, mold
	Product counterfeit	CountryA, supposed, CountryB, fan, merchants, legitimate, defect, Counter, fake, counter
	Product formulation issue	API, absence, guidelines, ethanol, curate, filters, too, varied, solvent, formulation
	Product label issue impacting strength, dose and/or safety	claimed, wheat, inadequate, matrix, claims, difference, absent, slim, word, Above
	Product mix up	mistake, CountryA, mix, Mix, mixing, shape, Prime, physical, short, discovery
Medium	Accompanying dose delivery device issue	fitted, failing, curled, inability, blunt, jammed, torn, transfer, attach, fails
	Advertisement non-compliance	permitted, website, claims, ables, infringement, producer, advertisement, CityA, permits, supplying
	Contamination with chemical substance	cloud, migration, laden, detection, contamination, Mercury, silicon, radioactive, explosives, metal
	Contamination with foreign matter	Matters, physical, invisible, rigid, traces, Chip, looks, lump, string, debris
	Lack of efficacy	doubtful, ineffective, eligible, declined, efficacy, lacked, erosion, announce, exhibits, concern
	Manufacturing non-compliance	Missing, installation, imprint, missing, External, residue, ABS, fault, connector, absence
	Out of specification or out of trend test result	project, submission, validity, ots, quantities, fines, exceeded, decrease, scent, sediment
	Product adulterated and/or contains prohibited substance	cone, western, hydrogen, authorised, rdi, Investigation, existence, Just, declared, poison
	Product deposit	sediments, erosion, dissolution, precipitation, crystal, hardened, thickness, deposits, dissolved, formation
	Product expiration date missing, illegible or incorrect	calendar, differing, expired, dates, date, periods, selecting, life, caps, life
	Product leaflet issue impacting strength, dose and/or safety	statements, Manual, claims, amendments, sticking, insert, Amendment, lets, holders, tape

	Product primary packaging issue	scratch, pier, detached, barrels, tear, leaked, positioning, hole, leaking, worn
	Product unregistered	discontinued, mistaken, tad, Groups, escape, certification, authorization, metro, commercial, Author
Low	Product label/leaflet issues NOT impacting strength, dose and/or safety	Manual, omitted, marking, address, packing, tape, matching, code, Editorial, elsewhere
	Product physical issue	shaved, design, freezing, Broken, squash, spots, oversized, Empty, yellowish, choking