Explainability for NLP in Pharmacovigilance: A Study on Adverse Event Report Triage in Swedish

Luise Dürlich[1,2,4] **Erik Bergman**[1] **Maria Larsson**[1]
Hercules Dalianis[5] **Seamus Doyle**[1] **Gabriel Westman**[*1,3] **Joakim Nivre**[*2]
[1]Swedish Medical Products Agency, Sweden
[2]Department of Linguistics and Philology, Uppsala University, Sweden
[3]Department of Medical Sciences, Uppsala University, Sweden
[4]RISE Research Institutes of Sweden, Sweden
[5]Department of Computer and Systems Sciences, Stockholm University, Sweden
luise.durlich@lakemedelsverket.se

Abstract

In fields like healthcare and pharmacovigilance, explainability has been raised as one way of approaching regulatory compliance with machine learning and automation. This paper explores two feature attribution methods to explain predictions of four different classifiers trained to assess the seriousness of adverse event reports. On a global level, differences between models and how well important features for serious predictions align with regulatory criteria for what constitutes serious adverse reactions are analysed. In addition, explanations of reports with incorrect predictions are manually explored to find systematic features explaining the misclassification. We find that while all models seemingly learn the importance of relevant concepts for adverse event report triage, the priority of these concepts varies from model to model and between explanation methods, and the analysis of misclassified reports indicates that reporting style may affect prediction outcomes.

1 Introduction

Pharmacovigilance (PV) deals with the detection, assessment, understanding and prevention of adverse effects related to medical products (World Health Organization, 2002) and traditionally relies on experts processing adverse event reports (AER), assessing the strength of new adverse event signals and acting upon newfound insights through publications and new risk assessments. In recent years, a need for at least partial automation has been identified to deal with the ever increasing amount of new AERs (Bate and Hobbinger, 2021) and at times updated processing requirements, most notable during the recent COVID-19 pandemic.

With the introduction of automated methods into the PV pipeline, experts have encouraged employing interpretable or at least explainable systems to address safety concerns such as black swan events (Kjoersvik and Bate, 2022) and including explainability as a factor to assess the readiness of artificial intelligence (AI) for tasks in the context of PV (Ball and Dal Pan, 2022). At the same time, concerns have been raised about the effectiveness of existing explainability methods and the disconnect between expectations towards explanations of black-box models from an AI safety perspective and what common explainability approaches actually are able to achieve (Ghassemi et al., 2021).

In this study, we apply two feature attribution methods to several pre-trained language models, fine-tuned to triage AERs, to understand what characterises their prediction of specific classes and to address the following research questions:

1. How do explanations for different models fine-tuned for the same task differ?
2. Can we align important features with regulatory criteria for serious cases?
3. Are there systematic feature patterns that explain incorrect class predictions?

Our analysis suggests that relevant features relating to regulatory criteria and expert annotation practice are learned as indicators of serious events by all models. However, the relative importance between these features in the explanations vary from model to model. Beyond features directly associated with serious reports, we find evidence of model bias reflecting the reporting style by different reporter groups.

2 Background

Explanations for machine learning models and their predictions come in many different forms. In light of model development and the paradigm shift to large generative models, several works have explored using large language models (LLMs) to explain their own output (Kunz et al., 2022; Kunz and

[*]Equal contribution to this work as senior authors.

Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), pages 46–68

Kuhlmann, 2024; Turpin et al., 2023). However, these works also warn that while such explanations may seem plausible to humans, it is unclear how well they represent the real reason for a specific model prediction, and Turpin et al. (2023) find evidence that they may in fact systematically misrepresent the deciding factors in the decision process.

Traditionally, deep learning models are often explained with so called post-hoc methods that are applied after the model is trained for a particular task. Methods such as diagnostic classifiers (Hupkes et al., 2018) are popular to answer specific questions about the encoded knowledge in a specific layer of the model by using representations of the chosen layer as input to a simpler model to perform a relevant task. More recently, Bricken et al. (2023) proposed the use of sparse auto-encoders to extract interpretable monosemantic features from single layer transformers. Templeton et al. (2024) applied this technique to the intermediate layer of smaller LLMs.

Feature attribution methods, such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017), instead attempt to explain model predictions by assigning some form of contribution to features in the input. These methods work by approximating the model to be explained on a given input using a more interpretable model, for example by perturbing the input in some way, observing the behaviour of the model to be explained, and explaining it with an explanation model trained to mimic that behaviour. Feature attribution methods can furthermore be model-agnostic, such as LIME and some versions of SHAP, or model-specific, such as gradient-based methods like DeepLift (Shrikumar et al., 2017) and Integrated Gradients (Sundararajan et al., 2017).

The feature attribution methods mentioned so far are typically applied to individual examples and thus primarily provide local explanations, but global explanations can be derived from local explanations by aggregating them over many inputs, e.g. using algorithms such as Submodular Pick LIME (Ribeiro et al., 2016) and Global Attribution Mapping (Ibrahim et al., 2019), or by simply averaging the observed attribution scores for each feature (Van Der Linden et al., 2019; Saynova et al., 2023).

Common goals for using explainability are model development, gaining trust, scientific insight and regulatory compliance (Hauben, 2022), but existing methods are criticised for suffering from interpretability gaps, failing to meet the expectations of stakeholders such as regulators and practitioners, and being prone to confirmation bias (Ghassemi et al., 2021). Moreover, Vilone and Longo (2021) note the absence of a common definition of explanations and lack of consensus on how to evaluate them with respect to reliability and validity. Further, while user-oriented explainability may be built with the intention of being simplified enough to be understandable, such explanations can be too far removed from the original model to faithfully represent it (Rudin, 2019).

Despite the concerns and criticisms toward post-hoc methods and feature attribution in particular, this type of explainability method is popular in natural language processing (NLP) research, where it has been used to achieve a variety of goals, such as providing insights into performance differences between different model architectures (Wang et al., 2022; Amponsah-Kaakyire et al., 2022), investigating potential weaknesses of explainability methods (Tang et al., 2022), interpreting aspects of the behaviour of pre-trained language models in specific NLP tasks (Nayak and Timmapathini, 2021; Stevens and Su, 2021), serving as reference explanations for investigating attention as an explanation method (Jain and Wallace, 2019), exploring descriptive features for distinct classes in domain-specific texts (Saynova et al., 2023), and user studies on computer-assisted coding tools (Dolk et al., 2022).

3 Method

Our experiments concern four binary classifiers fine-tuned on the same data for which we analyse post-hoc explanations derived with two types of feature attribution methods – Integrated Gradients (IG) (Sundararajan et al., 2017) and Expected Gradients (EG) (Erion et al., 2021). We restrict the study to these two gradient-based methods.

3.1 AER Triage

The classification task is that defined by Bergman et al. (2023): for AERs from both consumers and healthcare professionals, predict whether a report discusses a serious adverse reaction or not, based solely on free-text fields such as the adverse event terms listed in the form (e.g. headache, nausea, rash) and the description of adverse events in the report. An adverse reaction is considered serious if it results in death, is life-threatening, leads to hospitalisation or prolongs existing hospitalisation,

Dataset	Time period	Number of AERs			μ length
		S	NS	Total	
Training	2017 – 2020	4,450	7,538	11,988	$73.10_{\pm 70}$
Development	2017 – 2020	1,107	1,890	2,997	$70.30_{\pm 62}$
Test	2021 – 2021	1,170	2,273	3,443	$60.79_{\pm 68}$

Table 1: Overview of the three data sets used, with time periods, number of serious/non-serious (S/NS) reports and mean report length in whitespace-tokenised tokens.

Model	Abbreviation	Domain
KB-BERT	KBB	General
SweDeClin-BERT	SDCB	Clinical Text
AER-BERT	AERB	AER
GPT-SW3	GPT	General

Table 2: Selected models and their domains.

results in persistent or significant disability or incapacity or birth defects (ICH, 1994). When submitting an AER, reporters are asked to indicate these specific outcomes if they apply in a multiple-choice question. Replies to the question are among other things used to prioritise which reports get processed first by the case workers at the Swedish Medical Products Agency (MPA). However, the question is not always answered correctly given other context provided in the report, resulting in serious reports getting processed later than is desirable.

3.2 Data

The Swedish AERs that we base our training and explanation analysis on have been collected by the MPA and were annotated for seriousness by expert assessors as part of the agency's routine PV monitoring. We train the classifiers with the same training and development split as Bergman et al. (2023) and conduct a final evaluation of all four classifiers on the same prospective test set; see Table 1. Since we were able to obtain an improved version of the data used by Bergman et al. (2023), we conduct new hyperparameter experiments for all models described in the next section. Details on differences from the data used in (Bergman et al., 2023) and hyperparameter settings are in Appendix A. To remove numerical information related to identity, all reports were anonymised by replacing digits in the free-text description.

3.3 Models

We train four classifiers based on a selection of pre-trained transformer models for Swedish with various degrees of specialisation to the medical and AER domain. The first three are BERT models: the cased versions of KB-BERT (KBB) (Malmsten et al., 2020); SweDeClin-BERT (SDCB), a continuation of KB-BERT with additional pretraining on a corpus of de-identified clinical text (Vakili et al., 2022);[1] and AER-BERT (AERB), a masked-language model based on a large BERT model[2] with continued pretraining on old AERs. AER-BERT was previously found to give the best performance on the triage task by Bergman et al. (2023), compared to LSTMs and XGBoost models. In addition, we consider a small transformer decoder in the 1.3B parameter model of the GPT-SW3 model suite (GPT) (Ekgren et al., 2022). See Table 2 for an overview of the models.

We fine-tune all four models for the triage task by adding a classification layer to the pooled output of the transformer models using the applicable ForSequenceClassification classes implemented in the HuggingFace transformers library. Table 3 shows the classification performance of the four models on the test set. Among typical metrics for classification problems such as precision, recall and F_1, we also consider specificity, the true negative rate, to assess how well the models discriminate non-serious reports. We observe GPT to outperform all other models in F_1-score followed closely by AERB, and SDCB to perform best in specificity.

Model	Accuracy	Precision	Recall	Specificity	F_1
KBB	0.819	0.833	0.583	0.940	0.686
SDCB	0.813	0.891	0.512	0.967	0.650
AERB	0.830	0.845	0.612	0.943	0.710
GPT	0.822	0.788	0.653	0.909	0.714

Table 3: Classification results on the test set.

3.4 Feature Attribution Methods

This study considers two model-specific feature attribution methods, IG and EG. Both methods base their attribution on the notion of a *baseline* or *reference*, typically defined as a neutral or uninformative input for the task the model was trained for.

Integrated Gradients (IG): IG attributes the model prediction by calculating the path integral over gradients on a straight-line path from an artificial baseline input representation to that of the real

[1] Further research involving SweDeClin-BERT, like the training and analysis in this study, has been approved by the Swedish Ethical Review Authority under permission number 2022-02389-02.

[2] AI-Nordics/bert-large-swedish-cased

input. IG satisfies a number of desirable axioms for explainability methods as defined by Sundararajan et al. (2017), in particular sensitivity, implementation invariance, completeness, linearity and symmetry preservation, described in Appendix B.

Expected Gradients (EG): EG is a method inspired by IG that samples multiple real examples for reference and computes feature importance as the average expected values of the gradients scaled to satisfy the completeness axiom (Erion et al., 2021). Being gradient-based and symmetric, EG also fulfills the axioms defined for IG.

3.5 Explanation Methodology

To obtain explanations, we use the IntegratedGradient and GradientShap classes as implemented by the captum library (Kokhlikyan et al., 2020) for IG and EG, encoding all reports and baselines prior to applying the feature attribution methods. We compute feature attributions over the full encoder (or decoder) block and the classfication layer. For IG we create a report specific baseline consisting of a sequence of all `[MASK]` tokens for BERT models and <unk> for GPT, of the same length as the real report and pass along the attention mask for the real report to predict whether report and baseline are serious.[3] Each report is explained with 100 approximation steps. For EG we pass the entire set of reports in the development data as references. This way, each report is explained with respect to the ensemble of all other reports.[4] Here, we pass an extra argument containing report-specific attention masks.

With our binary classification task, explanations for serious and non-serious outcomes are symmetric in that large positive values explaining a serious prediction correspond to large negative attributions when explaining the opposite prediction for the same report. For consistency, all attribution values discussed in the following are computed with respect to predicting the serious class.

In the following experiments, explanations are obtained for 2,997 reports in the development set. When computing explanations for the four models, the explanation methods return results on token-level, i.e. referring to subwords as defined by the respective tokenizer. These representations are too fine-grained and hard to interpret and do not allow for easy comparison between models. To achieve a more global insight and allow for a more direct comparison between models, we calculate attributions at word-level as the sum of the corresponding token-level attributions per word. This is motivated by the axiom of completeness (Sundararajan et al., 2017), according to which the sum of attributions for an input sequence should reflect the difference in model prediction for the real input sequence and the baseline.

When reconstructing the vocabulary, the different tokenizers used by the models result in some slight variations in the complete sets of reconstructed word types, with 17,594 words according to KBB and SDCB, 17,585 for AERB and 17,612 with GPT.

To address the first two research questions, we compute global explanations on the development set reports for each model and feature attribution method using the normalisation method in Van Der Linden et al. (2019) and Saynova et al. (2023), effectively calculating global explanations as the relative attribution score for each full word in the dataset.

3.6 Analysing Explanations

Using global explanations for each classifier and explanation method, we want to analyse the attributions for interesting groups of related terms. To that end, we define the overall *importance* of each group as the average attribution value per model, and adjust for variation within the groups by scaling with the unbiased sample standard deviation:[5]

$$importance_g = \frac{\mu_g}{1 + \sigma_g} \quad (1)$$

In this way, we can focus on groups that consistently show large positive attribution values. To obtain groupings of terms, we consider an unsupervised approach in the form of clustering as well as the following explicit resources:

- **MeSH:** Medical Subject Headings (Lipscomb, 2000) is an ontology for indexing biomedical information by the National Library of Medicine.
- **Filter terms:** Terms and word segments cre-

[3] A common baseline for IG in NLP is that of a zero vector (Sundararajan et al., 2017) or empty string (corresponding to all `[PAD]` tokens for transformer models), but we argue that the mask and unknown tokens are a better choice, because the chosen models were not trained to attend to padding tokens during neither fine-tuning nor pre-training.

[4] Due to the number of reports we consider the effect of explaining the report by itself to be negligible.

[5] In the following, this equation is referenced when used to avoid confusion with importance as a general concept.

ated and used by assessors at the MPA, in the absence of the triage model (see Appendix C.1).

- **Criteria grouping:** Based on the criteria for a serious adverse reaction (cf. Section 3.1), we select a set of terms using MeSH and Swedish MeSH,[6] grouping them into general terms and terms relating to specific concepts within the five criteria (see Appendix C.2).

4 Results

4.1 Model Differences on a Global Scale

To compare explanations for different models, we calculate Kendall's τ correlation between the global attributions for the shared vocabulary by all models as well as for the set of terms matching the filter terms. As a frame of reference for the fine-tuned models, we also compare each classifier with its newly initialised, but not yet fine-tuned counterpart, and label that the *control*.

Correlations of attributions on all shared terms at the top of Figure 1 are weakly positive among all fine-tuned models, with slightly stronger correlations between the encoder models as opposed to encoders and GPT for IG. Interestingly, IG attributions for the two models with domain-specific pretraining have a lower correlation with each other than with the general domain KBB, and SDCB's correlation with KBB is slightly lower than that of KBB and AERB. By comparison, correlations among EG explanations are much weaker, with the strongest signals between KBB and the domain-specific models. For both IG and EG, correlations with the corresponding control models are close to zero, as would be expected for explanations of models unfamiliar with the triage task.

This correlation approach includes many terms with attributions close to zero for which comparison or correlation is uninformative. To focus on more relevant terms, we select terms matching the filter terms and calculate the correlations on this subset. The results at the bottom of Figure 1 show stronger correlations for both IG and EG. For IG, the trends between models are similar to those for the shared vocabulary, with an increased similarity between GPT and AERB. The correlations for EG are weaker between GPT and the other fine-tuned models and slightly stronger between KBB and the domain-specific models. Comparing both methods, correlations between control models and fine-tuned models are relatively stable for EG in both the larger and the more specific sets of terms, while they are stronger for IG in the latter setting.

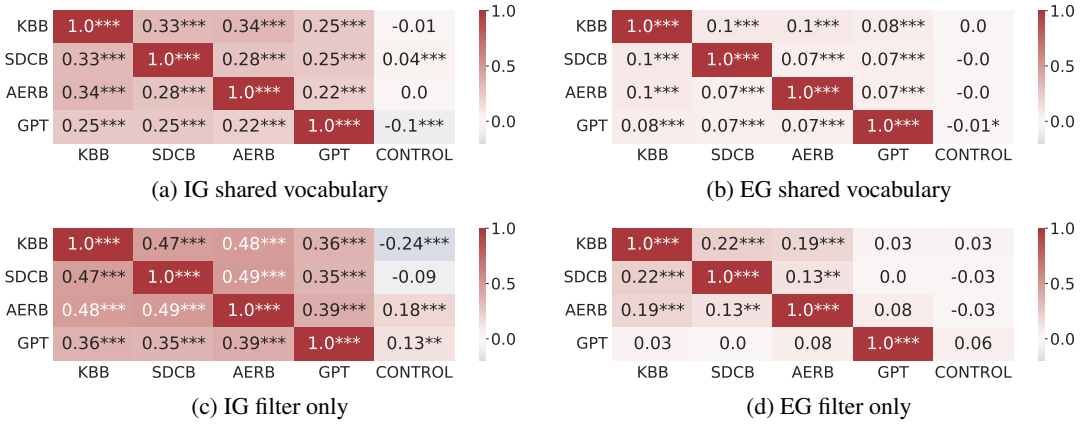
Based on the filter terms, we measure how highly the explainability methods score terms matching the filter, and the variance across models. Table 4 shows average attribution scores for three sets of terms: (1) words matching the filter, (2) words that do not match the filter, (3) all words in the dataset. Figure 6 in Appendix D visualises the distribution of scores in the first two sets for each model. All models trained for triage on average assign matched terms higher attribution scores than the ensemble of other terms. For the control models, all three sets have a similar average attribution score close to 0 for most models, suggesting no strong contribution to either the serious or the non-serious class for those terms. This indicates that all fine-tuned models learn to associate the filter terms with the positive class and that both explanation methods pick up on their importance.

Exploring more freely which concepts are important for a serious outcome with each model according to the explanations, we cluster terms with the largest attribution scores and hand-annotate the clusters. This resulted in 164 clusters for IG and 193 for EG, of which 134 had identical labels. We next consider how much of the clusters is covered by the 8,000 highest ranked terms and how *important* clusters are for each model as per Equation 1. Figures 2 and 7 show the twenty most important clusters to the average of all four models for IG and EG respectively. A two-dimensional visualisation of the full clustering reflecting cluster importance as explained by IG and EG can be found in Figures 13 and 14 in Appendix E, which also contains more details on the clustering procedure and coverage metric.

Considering explanations by IG, all classifiers note clusters relating to extreme situations (*suicide*, *ambulance*, *abortion*, *organ transplants*), organ-related issues, specific symptoms and health conditions (*depression*, *syncope*, *vision* and *breathing disorder*, *hypo-*,[7] *epilepsy*, *dementia*) as important. Importance by model varies somewhat, with *hallucination*, *breathing disorders* and *suicide* emerging as the most important clusters for KBB, while *ambulance* is less prominent. SDCB, in addition to *suicide* and *hallucination*, places more importance on

[6] https://mesh.kib.ki.se/

[7] Deficiencies denoted by terms with the prefix *hypo-*.



(a) IG shared vocabulary

(b) EG shared vocabulary

(c) IG filter only

(d) EG filter only

Figure 1: Kendall's τ correlations and their significance between models for shared vocabulary (a), (b), and filter terms (c), (d). The control row reports correlations, between each classifier and a corresponding untrained classifier.

(a) Fine-tuned models

	Model	In filter	Outside	All terms
IG	KBB	0.0348***	0.0008	0.0013
	SDCB	0.0634***	0.0095	0.0101
	AERB	0.0402***	0.0015	0.0020
	GPT	0.0699***	0.0103	0.0110
EG	KBB	0.0724***	0.0037	0.0046
	SDCB	0.0421***	0.0069	0.0073
	AERB	0.1000***	0.0062	0.0073
	GPT	0.0599***	0.0063	0.0069

(b) Control models

	Model	In filter	Outside	All terms
IG	KBB	-0.0044	-0.0032	-0.0032
	SDCB	0.0007	0.0001	0.0001
	AERB	0.0056***	0.0022	0.0023
	GPT	-0.0017	-0.0052	-0.0051
EG	KBB	-0.0009	-0.0004	-0.0004
	SDCB	-0.0032	0.0004	0.0003
	AERB	-0.0008	-0.0007	-0.0007
	GPT	0.0029	-0.0030	-0.0029

Table 4: Average attribution scores by explanation method for each of the four models. The scores are averaged for three sets of terms, those matching the filter terms, those not matching the filter terms and the report vocabulary as a whole. (a) shows results for the fine-tuned models and (b) shows results for the models prior to fine-tuning as a control. Significantly higher attribution scores of the filter terms compared those outside the filter are marked with * to *** to reflect the significance level of the Wilcoxon rank-sum test.

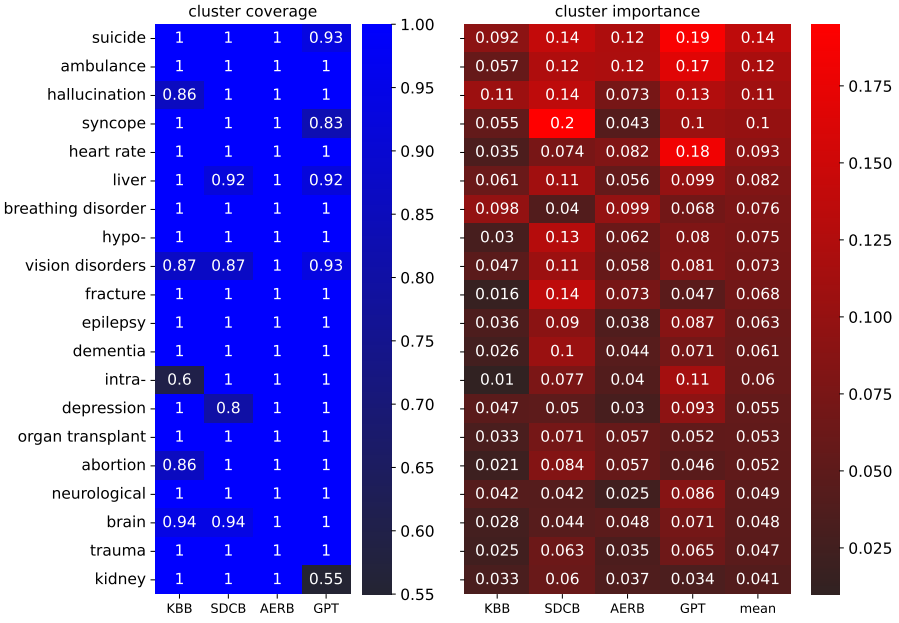
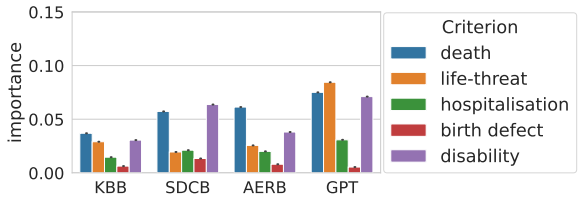
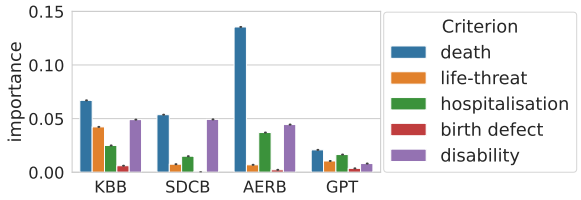


Figure 2: 20 highest ranked clusters by group importance (IG) and their coverage among the top 8,000 terms per model.



(a) IG



(b) EG

Figure 3: Group importance of different criteria for different classifiers and explanation methods.

the *syncope*, *fractures* and *hypo-* clusters. AERB is the only model with full coverage of all 20 clusters, but *hallucination* is less important, whereas *suicide*, *ambulance*, *breathing disorders* and *heart rate* are more important. Similarly, to GPT the most important clusters are *suicide*, *heart rate* and *ambulance*, but *hallucination* still ranks high.

An analysis of the EG explanations again reveals less overlap than IG among the most important clusters. However, we observe strong overlaps regarding cluster coverage among the top 3 clusters, those relating to symptoms as well as certain organ related issues. KBB is sensitive to specific events such as *suicide*, *childbirth*, *epilepsy*, but remains neutral on the *liver* and *abortion* clusters. SDCB only fully covers one cluster in the top 8,000 terms and along with *suicide* and *epilepsy* gives more importance to *liver*, *abortion*, *hallucination* and *hypo-*. For AERB, besides *suicide* and *liver*, *ambulance* emerges as most important and the *intra-* and *fainting* clusters receive more weight. Interestingly, among the domain-specific models, AERB assigns much more importance to *ambulance* than SDCB. To GPT, *hallucination* is most important, followed by *syncope*, *hypo-* and *blood*.

4.2 Regulatory Criteria

Figure 3 shows the *importance* of different criteria groups (see Appendix C.2) according to Equation 1. Overall, all criteria have a positive importance, indicating that the models learn their relevance without explicit exposure to the criteria. According to IG, death is one of the two most important ones for all models and disability is quite important in all four models. The life-threatening criterion appears most important with GPT, while it is much less important for the other models. In EG, death is the most important criterion for all models and disability is most important after that except for GPT, where hospitalisation is more important. With both methods, birth defect emerges as the least important criterion, but this may be because it is the smallest criteria group and infrequent in the data.

4.3 Analysis of Misclassified Examples

Preliminary analysis of misclassified examples revealed very few terms with deviant explanation patterns, which we took as an indication of issues with the gold labels of the AER data. As reported by Bergman et al. (2023), the annotation procedure of AERs at the MPA is suboptimal from a machine learning perspective, because of a regulatory guideline that assessors should not downgrade a report labelled serious by the reporter, even if they consider the report to contain no information meeting the criteria for serious events (EMA, 2017, p. 16). For this reason, we asked one of the assessors to reannotate all reports that were misclassified by both GPT and SDCB – the best models in terms of specificity and F_1, respectively – 345 reports in total. Appendix F gives statistics on the reannotated reports and shows that, for both false negatives and false positives, more than half of the labels changed, confirming our suspicions.

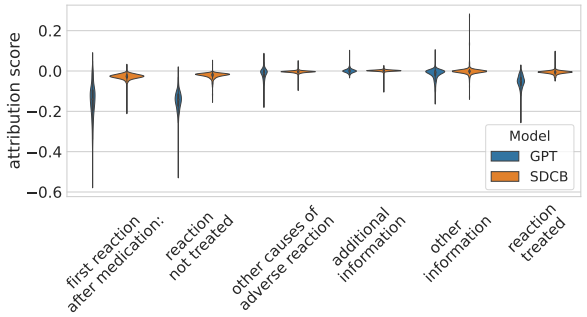
Given the new annotations, we identify the terms with the largest differences in attribution score between true and false predictions for both serious and non-serious reports, focusing on terms explained as more serious in either true positives (TP) vs. false negatives (FN) or true negatives (TN) vs. false positives (FP). Table 9 in the Appendix shows the terms matching the inclusion criteria, and Appendix G contains additional information on the selection of these terms. For both models we then separately consider local IG explanations of the reannotated reports containing these terms – about 130 reports per model – to see if there are systematic differences for TP/FN and TN/FP report pairs.

While the manual analysis guided by the terms did not reveal most of the terms themselves to have obvious systematic effects, we noted some trends observed over most of the reports with specific patterns often explained as more serious or non-serious than the average term. Investigating the usage of these patterns on the training set, we found evidence of them reflecting reporter groups and

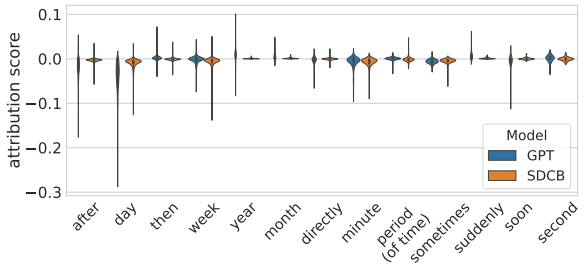
specifically stylistic differences in how consumers and healthcare workers report AERs. We found certain snippets of texts that occurred in many reports and that traced back to the original reporting form, which had several free-text fields that were then automatically concatenated and saved as one field with titles or generated text corresponding to specific answers. Such elements, referred to as form patterns in the following, were often explained as non-serious as a whole or in part. Another notable pattern was that of temporal references including mentions of periods of time (e.g. *minutes* or *days*), but also temporal adverbs like *soon* and *directly*, which were explained as non-serious by both models. Appendix H contains information about the specific patterns and their statistics on the training set. What these statistics illustrate is that most of the form patterns, with the exception of *other information:*, are almost exclusively used in consumer AERs. Although the reporting rates are less extreme for temporal patterns, terms like *sometimes*, *month* and *period* are more indicative of consumers, while *soon*, *minute* and *second* are slightly more used by healthcare workers.

We argue that some of the identified patterns align with how groups of reporters tend to express themselves in AERs, with healthcare personnel using medical jargon and writing concise reports,[8] while consumer reports can be longer and contain more detailed descriptions of how the reaction affected their everyday life and complaints about suspected products. From the form patterns we also observe that consumers appear to more diligently fill in the multiple free-text fields than healthcare workers who appear to rather give brief and to the point descriptions in one or a few of the fields.

In Figures 4a and 4b, we show how both types of patterns are explained by IG, plotting the distribution of their local explanations over the whole development set. Attribution scores were obtained by matching the exact sequence for form patterns, and summing the attribution scores of the individual words. Temporal patterns were matched with regular expressions covering morphologic variations.[9] In general, the explanations for SDCB appear more concentrated than those of GPT. Some form patterns like *first reaction after medication* and *reaction not treated* are clearly mostly negative in terms of attribution, i.e. explained as contributing to non-



(a) form patterns



(b) temporal patterns

Figure 4: Local attribution score distribution of form patterns and time references over all matched reports in the development set, ordered by frequency.

serious predictions, while *other information*, *additional information* and *other causes of adverse reaction* are more symmetrically concentrated around 0, suggesting an overall more neutral, less systematic contribution of these patterns to the prediction of reports in the development set. With respect to temporal references, there is a more global signal of *after*, *day*, *minute* and *sometimes* being explained as more non-serious with both models, while *directly*, *then* and *soon* appear slightly more neutral, and *year* and *suddenly* being explained as more serious.

The trends observed in attribution polarity and dominant reporting groups led us to take a closer look at model performance in these two groups in the development set. We found that recall for all four models was more than 20% lower for consumers than for healthcare workers and precision 10–20% lower. Correcting the gold labels where we have reannotations increases the scores for all models and subgroups, yet the differences in recall precision and F_1 persist for the subgroups of consumers and healthcare workers.[10]

[8] Although there is a variation with the type of profession, see the statistics in Table 12 in the Appendix.

[9] The expressions are listed in Table 11.

[10] More detail on this evaluation in Appendix J.

5 Discussion

The analyses in the previous section aimed at investigating feature attribution explanations for different triage models to answer the research questions defined in the beginning of the paper.

How do explanations for different models fine-tuned for the same task differ? To answer the first research question, we investigated the correlation between global attributions with two explainability methods. We found considerable variation between models, but also weak to moderate correlations among model attributions, most notably among encoder models and with the IG method. Moreover, models are more consistent with each other when task-relevant concepts are in focus as explored through filter terms and criteria groupings.

From the analysis of important clusters, we find that *suicide*, *ambulance* and *hallucination* appear in all models with both explanation methods. With IG, we can glean SDCB explanations to deem medical terminology such as *syncope*, *fracture* and deficiencies/dysregulations (*hypo-*) most important, while KBB, AERB and GPT focus more on the concepts common to all models, although AERB and GPT also give high importance to *heart rate*. With EG, we find some similarity in the most important clusters, with SDCB still having high importance scores for deficiencies, but also featuring other concepts like *epilepsy*, *abortion* and *liver*, while GPT retains *hallucination* as an important cluster, in addition to *syncope*, deficiencies and *blood*.

Can we align important features with regulatory criteria for what constitutes a serious case? All models seem to learn the importance of the filter terms and the groupings of criteria, albeit with different priorities as suggested by both the correlations over filter terms and the importance assigned to different criteria.

Are there systematic feature patterns that explain incorrect class predictions? Through the manual analysis of reports we learned that serious and non-serious explanations do not always focus on parts of the report that could be considered relevant for the assessment of the report at hand, and that the level of detail may be a factor contributing to misclassification. This raises the question whether the selected methods are adequate given the classification problem at hand and how one can conceptualise the two classes to distinguish. Is a non-serious report a distinct category in itself with salient features identifying it or just defined by the lack of serious features? And should we define an abstract neutral baseline or model explanations in contrast to the non-serious class?

6 Conclusion

In conclusion, our analysis shows that all models learn to identify relevant features indicative of a potentially serious case, but with varying focus on symptoms, conditions and medical procedures. Most of the criteria for identifying serious events are important for serious predictions with all models and explanation methods, but their relative importance varies across models. Finally, manual analysis of reports reveals features reflecting the reporting style of specific reporter groups, specifically reflecting which and how many free-text fields were filled in and to some degree the narration style and level of detail as represented through temporal references. This part of the analysis raises questions about model training and the adequacy of the selected explanation methods for the task at hand. Future work on training and explaining triage systems may need to rethink how information in this binary setup is defined and contrasted, to promote the importance of medically relevant features over confounding features related to form and writing style.

Limitations

In the preparation of this study, we made several design decisions that can be scrutinised further. In particular, the chosen explainability methods come with their own set of limitations, one of which is that, while feature attribution may highlight important terms, such a representation ultimately does not explain why the model that is being explained relies on those features to begin with. In addition, feature attribution for the most part constrains us to individual explanations of the input features without representing how features may interact with or affect each other. At the same time, the goal of the study in question was not to identify the best explanation technique for our use case, but instead to investigate triage models with available feature attribution methods.

We chose to focus on real-world data and models that may be employed as part of the MPA's pharmacovigilance monitoring. As such, the main focus of this paper was not to make claims on exact classification performance differences of the triage

models we analyse and we therefore did not pursue evaluation over several training seeds as this would also further complicate the analysis of explanations taking into account several versions of each fine-tuned model. For an analysis of the robustness of fine-tuning the AER-BERT model for triage we refer to our previous results in Bergman et al. (2023).

We did not study the effect of different fine-tuning runs on the final explanations given the same hyperparameters and base model and therefore cannot make any claims on how much of the differences we see between triage models is due to initialisation of the classification head, shuffling of the training data, or the difference in pre-trained base model. However, a limited control experiment showed that global explanations of ten fine-tuned versions of KB-BERT with different random seeds correlated much more strongly with each other than with any of the other models, which suggests that the differences between different pre-trained models are relatively robust. See Appendix K for more information.

The decision to use generative models with fine-tuning methods geared towards encoders instead of reframing the task into a generative setup may not have been the optimal choice for the GPT-SW3 model, but was chosen to follow a common methodology in deriving explanations and, most notably, always having a binary classification outcome space to refer to.

A large part of the analysis rests on aggregated attribution values. Corpus-level normalisation is only one way of achieving this aggregation. Furthermore, aggregation of explanations over multiple reports comes at the cost of losing nuances in specific contextualised cases.

Throughout the analysis, we consider raw aggregated values for each model. Using such unnormalised average attribution values means that global explanations between models are not directly comparable, since some models have much more extreme attribution values – this is why we took more of a ranking approach and focused on relative importance among, e.g., criteria groups.

The grouping of criteria is debatable for certain terms that may fit multiple categories or can be hard to disentangle in relation to another category (e.g. miscarriage as death rather than birth defect, cardiac arrest as death vs. life-threatening). Further, the groups are likely not an exhaustive list of relevant criteria terms in the given data, and as raised in the analysis, some groups cover only very few and overall infrequent terms and may provide a limited representation of the criterion in question.

Likewise, while the clustering analysis underwent several iterations to find a good separation of clusters without generating too many outliers there may be parameters resulting in an even better clustering result. In addition, to save resources, the clusters used in the analysis were manually labelled by a single annotator, based on the MeSH ontology and no further quality checks were conducted on this annotation. Involving more and more expert annotators in the process may have resulted in higher quality labels and slightly different grouping decisions for similar clusters and consequently different results. This could for example lead to combining more semantically similar clusters that are only distinguished by their level of specialisation such as the *fainting* and *syncope* clusters.

As for the investigation of reporter groups inspired by the manual analysis of explanations, one obvious aspect potentially dividing reporter groups is medical terminology and frequently used abbreviations by medical workers. While both references to medical conditions and procedures as well as drug names were noted as salient in some of the manually analysed reports, the variation of terms was larger and an exhaustive list more challenging to put together and analyse than the patterns we decided to study further.

Ethical Considerations

The data used in this work contains sensitive medical information and has been collected and processed by the Swedish Medical Products Agency as part of their pharmacovigilance monitoring duty. For the scope of this study, processing the data by training and evaluating models and their explanations falls under the agency's operations for business development and does not require further ethics approval by the Swedish Ethical Review Authority. To ensure information security, the texts have been anonymised by replacing digits in the free-text, where personal identity numbers may be reported. Further, complete examples of individual AER descriptions cannot be included without additional anonymisation steps. Since the study itself focuses on the explanation and evaluation of triage models for larger sets of reports this has not been necessary and observations are reported as summaries of subsets of the full AER data.

Acknowledgments

We acknowledge the support of the Swedish Research Council (grant no. 2022-02909). We also thank the reviewers for their helpful feedback and suggestions.

References

Kwabena Amponsah-Kaakyire, Daria Pylypenko, Josef Genabith, and Cristina España-Bonet. 2022. Explaining translationese: why are neural classifiers better and what do they learn? In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 281–296, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Robert Ball and Gerald Dal Pan. 2022. "Artificial intelligence" for pharmacovigilance: Ready for prime time? *Drug Safety*, 45:429–438.

Andrew Bate and Steve F Hobbinger. 2021. Artificial Intelligence, real-world automation and the safety of medicines. *Drug Safety*, 44:125–132.

Erik Bergman, Luise Dürlich, Veronica Arthurson, Anders Sundström, Maria Larsson, Shamima Bhuiyan, Andreas Jakobsson, and Gabriel Westman. 2023. BERT based natural language processing for triage of adverse drug reaction reports shows close to human-level performance. *PLOS Digital Health*, 2(12).

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L Turner, Cem Anil, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chirs Olah. 2023. Towards monosemanticity: Decomposing language models with dictionary learning, Transformer Circuits Thread, released 4 October 2023 [online].

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg.

Alexander Dolk, Hjalmar Davidsen, Hercules Dalianis, and Thomas Vakili. 2022. Evaluation of LIME and SHAP in explaining automatic ICD-10 classifications of Swedish gastrointestinal discharge summaries. In *Scandinavian Conference on Health Informatics*, pages 166–173.

Ariel Ekgren, Amaru Cuba Gyllensten, Evangelia Gogoulou, Alice Heiman, Severine Verlinden, Joey Öhman, Fredrik Carlsson, and Magnus Sahlgren. 2022. Lessons learned from GPT-SW3: Building the first large-scale generative language model for Swedish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3509–3518, Marseille, France. European Language Resources Association.

Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg, and Su-In Lee. 2021. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature Machine Intelligence*, 3(7):620–631.

European Medicines Agency. 2017. Guideline on good pharmacovigilance practices (gvp) - Module VI – Collection, management and submission of reports of suspected adverse reactions to medicinal products (Rev. 2). https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/guideline-good-pharmacovigilance-practices-gvp-module-vi-collection-management-and-submission-reports-suspected-adverse-reactions-medicinal-products-rev-2_en.pdf.

Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. 2021. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745–e750.

Manfred Hauben. 2022. Artificial intelligence in pharmacovigilance: Do we need explainability? *Pharmacoepidemiology and Drug Safety*, 31(12):1311–1316.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Mark Ibrahim, Melissa Louie, Ceena Modarres, and John Paisley. 2019. Global explanations of neural networks: Mapping the landscape of predictions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 279–287, New York, NY, USA. Association for Computing Machinery.

ICH Harmonised Tripartite Guideline. 1994. Clinical Safety Data Management: Definitions and Standards for Expedited Reporting E2A. In *International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use*.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. In *Proceedings of NAACL-HLT*, pages 3543–3556.

Oeystein Kjoersvik and Andrew Bate. 2022. Black swan events and intelligent automation for routine safety surveillance. *Drug Safety*, 45:419–427.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. 2020. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*.

Jenny Kunz, Martin Jirenius, Oskar Holmström, and Marco Kuhlmann. 2022. Human ratings do not reflect downstream utility: A study of free-text explanations for model predictions. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 164–177.

Jenny Kunz and Marco Kuhlmann. 2024. Properties and challenges of LLM-generated explanations. In *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 13–27, Mexico City, Mexico. Association for Computational Linguistics.

Carolyn E Lipscomb. 2000. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the national library of Sweden–making a swedish BERT. *arXiv preprint arXiv:2007.01658*.

Anmol Nayak and Hari Prasad Timmapathini. 2021. Using integrated gradients and constituency parse trees to explain linguistic acceptability learnt by BERT. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 80–85, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Denitsa Saynova, Bastiaan Bruinsma, Moa Johansson, and Richard Johansson. 2023. Class explanations: the role of domain-specific content and stop words. In *The 24rd Nordic Conference on Computational Linguistics*, pages 103–112.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMlR.

Samuel Stevens and Yu Su. 2021. An investigation of language model interpretability via sentence editing. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 435–446, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR.

Ruixuan Tang, Hanjie Chen, and Yangfeng Ji. 2022. Identifying the source of vulnerability in explanation discrepancy: A case study in neural text classification. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 356–370, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Brian Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C Daniel Freeman, Theodore R Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. 2024. Scaling monosemanticity: Extracting interpretable features from Claude 3 sonnet, Transformer Circuits Thread, released 21 May 2024 [online].

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pages 74952–74965. Curran Associates, Inc.

Thomas Vakili, Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. 2022. Downstream Task Performance of BERT Models Pre-Trained Using Automatically De-Identified Clinical Data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4245–4252.

Ilse Van Der Linden, Hinda Haned, and Evangelos Kanoulas. 2019. Global aggregations of local explanations for black box models. *arXiv preprint arXiv:1907.03039*.

Giulia Vilone and Luca Longo. 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106.

Jin Wang, Liang-Chih Yu, and Xuejie Zhang. 2022. Explainable detection of adverse drug reaction with imbalanced data distribution. *PLoS computational biology*, 18(6).

World Health Organization. 2002. *The importance of pharmacovigilance*. World Health Organization, Geneva. ISBN: 9241590157.

A Improved Data and Hyperparameter Experiments

The free-text description of reports in the original data used by Bergman et al. (2023) occasionally contained comments by the assessors processing

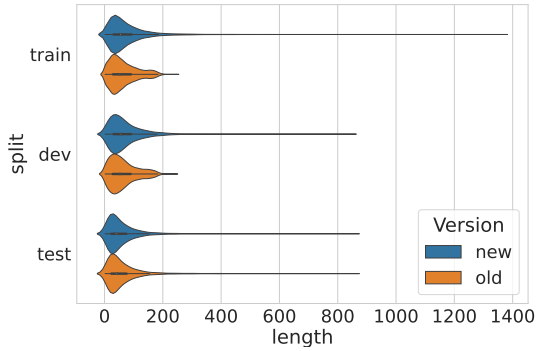


Figure 5: Report length in whitespace-tokenised tokens for the cleaner version of the data used in this paper (new) and the version previously used in Bergman et al. (2023) (old).

the report. In their work, preprocessing included filtering out and removing those comments using regular expressions. However, for this study we were able to obtain access to a database storing only the original reports as they were at the time of reporting and therefore skip this step in preprocessing the text. Upon comparing matching reports in the two data sources, we also discovered that the previously used data source contained truncated reports. Figure 5 shows a comparison of report lengths in the previous and current version of the data.

The database we extracted our reports from only contained those reports received by the MPA via an electronic reporting form. We found that some reports in the dataset used by Bergman et al. (2023) were not present in the original database and such cases could be explained by the original incoming reports covering information warranting a separate report, e.g. when the report describes adverse events related to different medical products at different points in time, specifically assigns different suspected events to different medication, mentions multiple patients with similar adverse events, or discusses events in mothers or soon-to-be mothers as well as events in their young children or fetuses. These reports were then split manually by assessors and added to the working database. Our data splits contain 90 such examples in development and training set, 42 of which were found to start with comments during pre-processing. To allow for some degree of comparison with our previous study, we opt to still keep these reports in their previous form and apply filtering to remove initial comments matching specific keywords followed by dates and assessor signatures.

Preprocessing for all reports includes stripping of initial hyphen characters and white space in the description field as well as prepending to the description all suspected adverse events in list form.

The focus of the hyperparameter experiments was to identify learning rate and epoch settings for the four models. We considered learning rates in the set {0.00002, 0.00003, 0.00004, 0.00005} and training for up to three epochs and chose the best settings according to the observed loss on the development set. Table 5 shows the selected settings informed by the experiments.

The settings for KBB and SDCB are identical. For AERB, we add a weight decay term of 0.01 to keep consistency with Bergman et al. (2023).

B Axioms of IG and EG

As defined by Sundararajan et al. (2017), the axioms fulfilled by both explanation methods are

- sensitivity, whereby only relevant features contribute to the explanation and irrelevant features have an importance of 0,
- implementation invariance, stating that for two networks that produce the same outputs as each other for all inputs, the attributions should be identical,
- completeness, in the sense that the sum of attributions for a particular input should correspond to the difference in model output for the input and the baseline,
- linearity, in that attributions for a model that is a linear combination of two other models are a linear combination of the attributions for those two models,
- symmetry-preservation, whereby symmetric variables in the network should get the same attribution if they have the same value.

C Analysis resources

C.1 Filter Terms

The list of filter terms contains 47 terms or segments that relate to words associated with serious reports and is used to filter incoming reports marked as not serious for candidates that can be prioritised. A drawback of its format is that word segments, not always representing real morphemes, may also match less relevant terms. All filter terms and approximate translations with annotations for

Parameter	KBB & SDCB	AERB	GPT
Batch Size	8	8	4
Gradient Accumulation	1	1	2
Learning Rate	2×10^{-5}	2×10^{-5}	2×10^{-5}
WarmupRatio	0.3	0.3	0.15
Mixed Precision	–	–	fp16
Optimizer	AdamW	AdamW	AdaFactor
Weight Decay	0	0.01	0
Epochs	1	1	2

Table 5: Training Settings

omitted parts are listed in Table 6. The filter terms match a total of 220 terms of the vocabulary in the global explanations.

C.2 Criteria Groups

The criteria groups are 5 groups of concepts derived from the definition of serious adverse reactions – relating to death, life-threatening reactions, hospitalisation, disability and birth defects. Each group consists of single word synonyms as well as more specific concepts, and is internally grouped to reflect more general notions as well as very specific terminologies and contexts.

For example, the group for death comprises a group of general words such as *death*, *pass away*, *passing* as well as individual groups for more specific forms of death such as *suicide*, *suffocation/asphyxia*, *cardiac arrest* and *miscarriage*. This grouping was created for the set of terms covered in the development set and is not exhaustive with respect to all possible subcategories that may exist outside this restricted vocabulary. Terms cover different wordforms of the same lexeme.

Table 7 shows how many terms and subgroups are associated with each criterion. The biggest criterion is that of hospitalisation with 179 terms. These include different inflected versions of the same lemma as well as common abbreviations and in some cases spelling variations found in the corpus of AERs that constitute the development set. The groups were created using MeSH and referring to terms present in the AER reports. Hence some groups such as birth defect are fairly small even though there are more conceivable birth defects, but they do not feature in the analysed set of AERs.

Filter Term	Translation
ARDS	respiratory distress syndrome
BNP	brain natriuretic peptide
Haemoly	haemoly(sis)
Johnson	Johnson
andningsavbrott	respiratory arrest
andningspåverkad	respiratory challenged
andningssvikt	respiratory failure
andningsuppehåll	respiratory arrest
anfall	attack, acute onset
avled	died
barre	Barre (Guillain-Barré syndrome)
blind	blind
cerebro	cerebro-
dog	died
dyspne	dyspnea
död	death
epidermal	epidermal
epilep	epilep(sy)
fladder	flutter
hallucin	hallucin(ation)
handik	disab(ility)
hemolyti	hemolyti(c)
hörsel	hearing
interstit	interstit(ial)
kardiell myopati	cardiomyopathy
koagulat	coagulat(tion related)
kolangit	cholangitis
konstaterad	confirmed / diagnosed
lungsvikt	lung failure
lymphohist	lymphohist-
mikroangio	microangio-
missbild	malforma(tion) / birth defect
missfall	miscarriage
multisystemisk	multisystemic
mungip	corner of the mouth
optikusneu	optic neu(ritis)
propp	clot
puls	pulse
purpura	purpura
resp insuff	resp(iratory) insuff(iciency)
scars	scars
syn	vision
synbortf	(loss) of vision
toxisk	toxic
vaerd	vaccine-associated enhanced respiratory disease
ventrike	ventric(le)
ventrombos	venous thrombosis

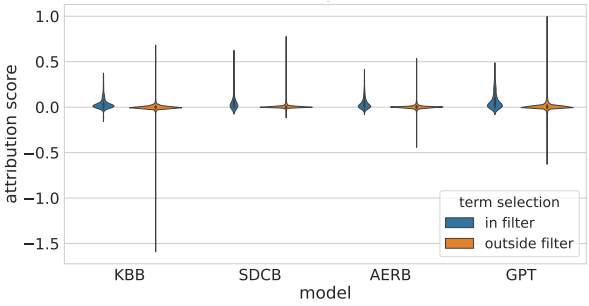
Table 6: 47 Swedish filter terms and their English translations and completions.

Group	Terms	Subgroups
Death	33	5
Life-threatening	10	1
Hospitalisation	179	3
Birth defect	4	2
Disability	20	5

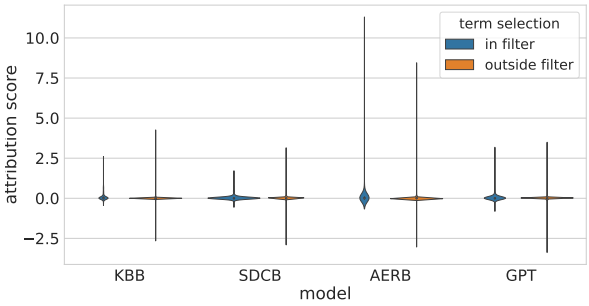
Table 7: Total number of terms and subgroups in each of the criteria groups.

D Feature Attribution for Filter Terms and Non-Filter Terms

Figure 6 shows the distributions of global attribution scores for terms matching the filter and those not matching the filter with both IG and EG.



(a) IG



(b) EG

Figure 6: Distribution of global attribution scores for terms matching the filter and terms not matching the filter.

E Clusters of Top 8000 Serious Terms

To find more general concepts important for a serious outcome with each of the models according to either explanation method, we took the union of the 8,000 most important terms per model and clustered them for each attribution method. Terms were first embedded using a Swedish Sentence-BERT model[11] and then decomposed to 50 dimensions using principal component analysis with whitening and clustered with HDBSCAN (Campello et al., 2013). We experimented with lemmatization at an earlier stage, but found it harder to obtain an interpretable clustering that way. We set the HDBSCAN clusterer to a maximum cluster size of 80, a minimum cluster size of 5 and used default settings for the remaining parameters. The clusters were annotated by hand by a single annotator with a background in linguistics and good command of Swedish. To make sense of medical terminology and how medical concepts relate to each other, the annotator relied heavily on MeSH and its Swedish version to derive sensible cluster names in English. Table 8 shows statistics on the amount of selected terms per feature attribution method, the number of resulting clusters, average cluster sizes and the amount of outliers.

Figure 7 shows the importance of clusters in EG and to what extent they were covered by each model's top 8,000 terms. **Coverage** in the latter visualisation refers to the number of terms belonging to the cluster, that also rank among the top 8,000 terms for a particular model, divided by the total size of the cluster in unique terms.

Figures 13 and 14 show the entire clustering of IG and EG reduced with t-SNE. For both IG and EG, some clusters are completely missing in the global explanations of certain models, due to different tokenization. Specifically, AERB and GPT pick up certain *units* (*µg*, *µmol*) that are missing for KBB and SDCB, and all models but GPT pick up numbers and dimensions describing affected areas listed as part of the adverse event terms, because GPT's tokenizer splits them into digits belonging to a separate cluster instead.

F Reannotation

Figure 8 shows how the FN and FP reports were annotated by the assessor given only the concatenated term list and description text field. We anticipated that annotating these without the usual context may complicate decision making for the assessor and therefore allowed both an unclear annotation and a field to comment on the annotation. For the entire 345 reports, only 7 cases were unclear without additional information.

Looking at the label proportions, out of the serious reports in the original gold annotation, predicted non-serious by both models (FN), only a third was actually serious after the reannotation. Of the reports originally annotated non-serious, but

[11] KBLab/sentence-bert-swedish-cased

Method	Terms in Union	Clusters	Terms per Cluster	Outliers
IG	13,909	164	8.3	12,547
EG	15,347	193	8.4	13,726

Table 8: Statistics on the clustering.

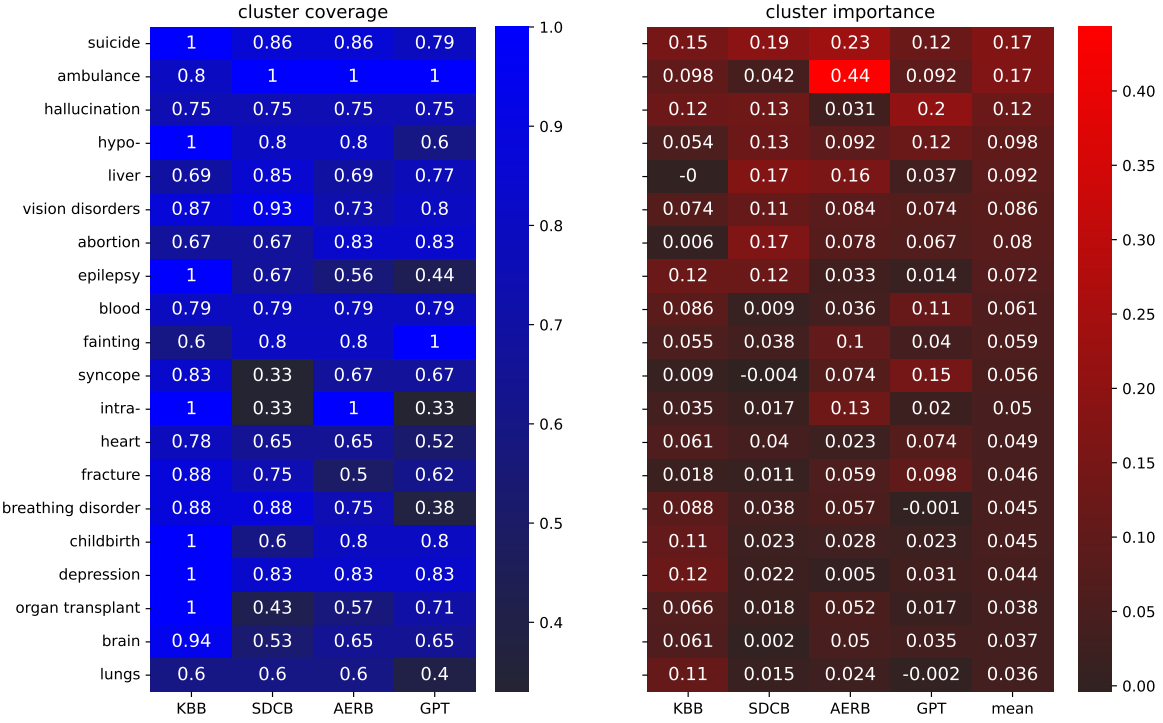


Figure 7: 20 highest ranked clusters with EG by cluster importance (right) and their coverage among the top 8,000 terms per model.

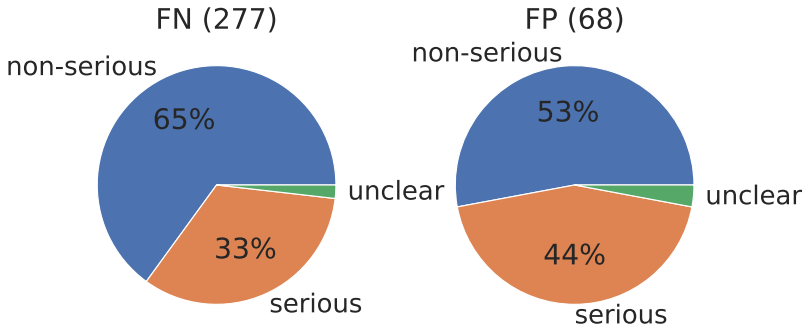


Figure 8: Reannotation of False Negatives (FN) and False Positives (FP). The numbers in parentheses are the amount of reports in each category.

predicted serious, about half remained non-serious after reannotation. One possible reason for the label change of so many of the originally FP reports is that some context is omitted with respect to the original report, since AERs consist of more than just the term list and free-text and the information indicating a serious event could conceivably be other parts of the form or its attachments without it being mentioned in the text as seen by the model.

G Selecting Reports for Manual Analysis

To identify interesting reports in the set of reannotated reports, we compute the terms with the largest differences in attribution score between true and false predictions for both serious and non-serious reports and restrict this to the 5 most extreme terms that occur at least twice in each considered set of reports with differences in the 2.5- and 97.5-percentiles respectively.

To limit the scope of the manual analysis, we only do this calculation and the report-wise analysis with IG. Table 9 details the terms, and their translation for the contrasted sets and each model.

The terms comprise some reoccurring themes for both models with terms relating to specific events such as *vaccination* or *product exchange*,[12] references to respiration (*breathing*, *coughing* and *shortness of breath*), the *emergency room*, and the abbreviation *EVF* for a blood test measuring the volume of packed red blood cells in a sample. They match a total of 126 and 129 reports for SDCB and GPT respectively. For each report we summarise the text and take note of the terms explained as serious and non-serious using IG as well as whether they relate to the specific event, fall under additional information such as patient history or information on other people mentioned in the report, or are stylistic elements of the report.

Analysing the reports associated with most of the terms in Table 9 revealed a variety in cases and narratives, however, there was overlap between the matched reports for *vaccination*, *vertigo*, *nausea* and *swelling* frequently co-occurring.

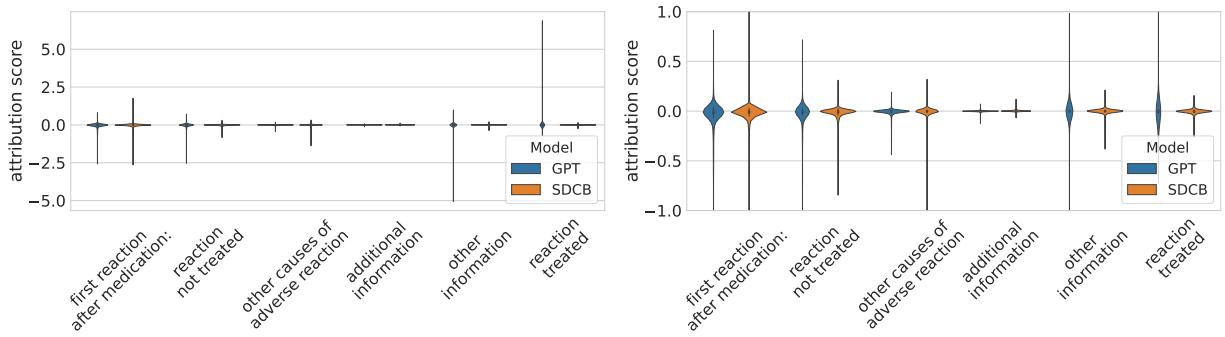
H Patterns

Table 10 details the six form patterns identified during the manual analysis. They correspond to automatically inserted titles or text snippets expressing information like whether or not the suspected adverse reaction was treated or how long after the affected person took the medicine suspected of causing the AE they started experiencing symptoms.

Table 11 details the Swedish temporal references as regular expressions to cover morphologic variation such as singular and plural, and indefinite and definite forms for nouns, and synonyms or contractions of some of the adverbs, with English translations and statistics on the occurrence of these terms in the training set and how much of those are in consumer reports.

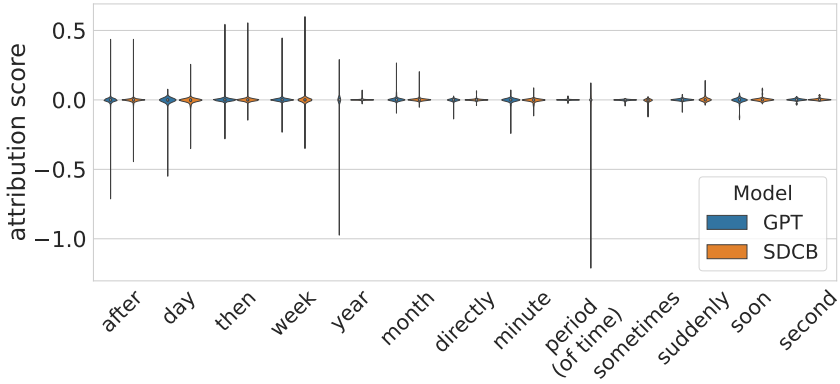
Figure 9 shows the attribution distributions of form and temporal patterns according to EG, which generally appear to be explained as more neutral than those by IG.

[12] Referring to cases when the intended prescribed product is replaced by an equivalent product by another pharmaceutical company, which can happen when the intended product is out of stock at a pharmacy.



(a) form patterns

(b) form patterns (zoomed in)



(c) temporal patterns

Figure 9: EG attribution scores of form patterns and temporal references in the full development set. The patterns are ordered by frequency in the development set with the most frequent patterns to the left.

Contrasted	**SDCB**		**GPT**	
sets	FN more serious	TP more serious	FN more serious	TP more serious
TP & FN	produktutbyte, andas, hosta, vaccination, rygg	blod	akuten, andfåddhet, smärtor, blod, andas	hosta, produktutbyte, biverkan, yrsel, reaktionen
English	*product exchange, to breathe, cough / to cough, vaccination, back*	*blood*	*(the) ER, shortness of breath, pains, blood, to breathe*	*cough / to cough, product exchange, (the) adverse reaction, vertigo, (the) reaction*
	FP more serious	TN more serious	FP more serious	TN more serious
TN & FP	akut, stroke, syn, svullna, evf	klåda, akuten, biverkningsombud, rodnad, dagar	evf, yr, migrän, yrsel, torra	stroke, syn, akuten, akut, EVF
English	*acute / ER, stroke, vision, swollen, packed red-cell volume*	*itching, ER, AER-delegate,[13] redness, days*	*packed red-cell volume, nauseous, migraine, vertigo, dry*	*stroke, vision, (the) ER, acute / ER, packed red-cell volume*

Table 9: Terms with more extreme differences in attribution score in correct and incorrect predictions per report class.

Pattern	Translation	Occurrence	Reported by Consumers
första reaktionen efter medicineringen:	first reaction after medication:	5,173	99.65%
reaktionen ej behandlad	reaction not treated	3,853	99.77%
andra biverkningsorsaker:	other causes of adverse reaction:	3,433	99.65%
ytterligare info	additional information	2,123	99.06%
övrig information:	other information:	1,903	0.08%
reaktionen behandlad	reaction treated	1,591	99.43%

Table 10: Swedish form patterns, their English translation, occurrence in the training set and the proportion reported by consumers.

Pattern	Translation	Occurrence	Reported by Consumers
(där)?efter	after	8,481	63.12%
dag(en|ar(na)?)?	(the) day, (the) days	3,990	63.73%
se(da)?n	then	2,654	66.11%
veck(an?|or(na)?)	(the) week, (the) weeks	2,020	61.49%
år(et|en)?	(the) year, (the) years	1,373	69.56%
månad(en|er(na)?)?	(the) month, (the) months	1,382	74.75%
direkt	directly	658	66.11%
minut(en|er(na)?)?	(the) minute, (the) minutes	449	46.55%
period(en|er(na)?)?	(the) period (of time), (the) periods	260	74.62%
ibland	sometimes	319	88.71%
plötsligt	suddenly	198	69.19%
strax	soon	147	46.26%
sekund(en|er(na)?)?	(the) second, (the) seconds	73	49.32%

Table 11: Regular expressions for temporal patterns in Swedish, their English translation, occurrence in the training set and proportion reported by consumers.

I Reporter Statistics

Table 12 contains statistics on reports by specific reporter groups in the training data.

Reporter	**Number of reports**	**Average report length** (in characters)
Consumer	5,607	614.04
Doctor	3,687	408.15
Nurse	1,573	364.00
Pharmacist	955	281.31
Dentist	131	301.58
Other Healthcare personnel	35	869.31
All Healthcare	6,381	378.63

Table 12: Statistics by reporter group on the training set

J Subgroup Performance

Figures 11 and 12 show the performance of each model in different metrics for the original development set and partially corrected gold labels.

[13] A delegated nurse / pharmacist reporting adverse events from the medical record system on behalf of a hospital.

K Explanation Correlation with Different Fine-Tuning Runs of the Same Model

This section shows results of a control experiment comparing global correlations for different fine-tuned versions of the same base model with the results in Section 4.1.

	Shared vocab.		Filter terms	
Base model	IG	EG	IG	EG
KBB	$0.65_{\pm 0.06}$	$0.17_{\pm 0.01}$	$0.64_{\pm 0.07}$	$0.20_{\pm 0.04}$
Different	$0.28_{\pm 0.05}$	$0.08_{\pm 0.01}$	$0.42_{\pm 0.06}$	$0.11_{\pm 0.09}$

Table 13: Average Kendall's τ correlation between explanations of 10 different fine-tuning runs of KBB and the different base models as reported in Figure 1 (excluding controls and the diagonal).

We fine-tuned 10 versions of KBB with the same hyperparameter settings as the model reported in the main text, but different random seeds to observe how similar global explanations are with the same pre-trained model. Table 13 shows average Kendall's τ correlations and their standard deviations for explanations of these new fine-tuned models sharing the same base model and the corresponding values for the experiments with different fine-tuned base models from Figure 1.

Figure 10 gives a better view of the distribution of these correlations

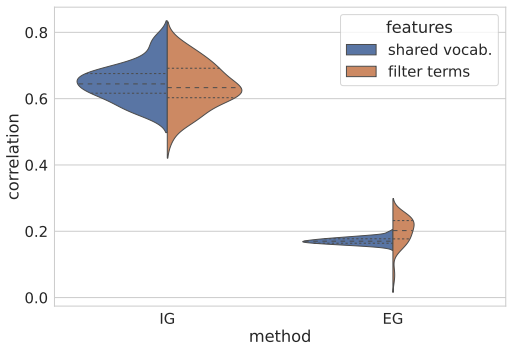


Figure 10: Distribution of Kendall's τ correlation between global explanations of 10 different fine-tuned KBB models.

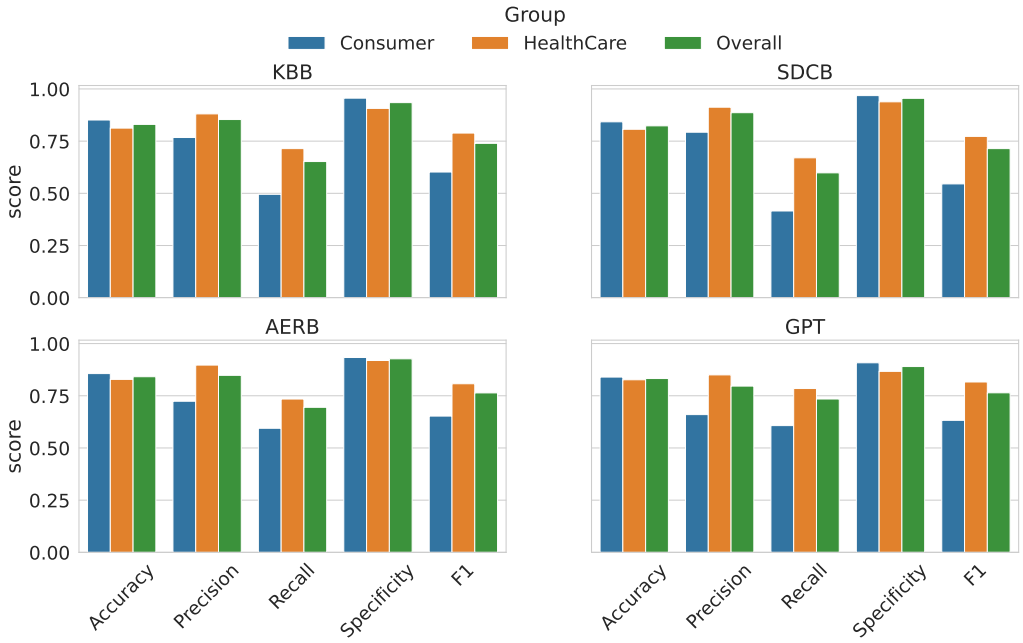


Figure 11: Model results on development data for reporter subgroups on original gold labels.

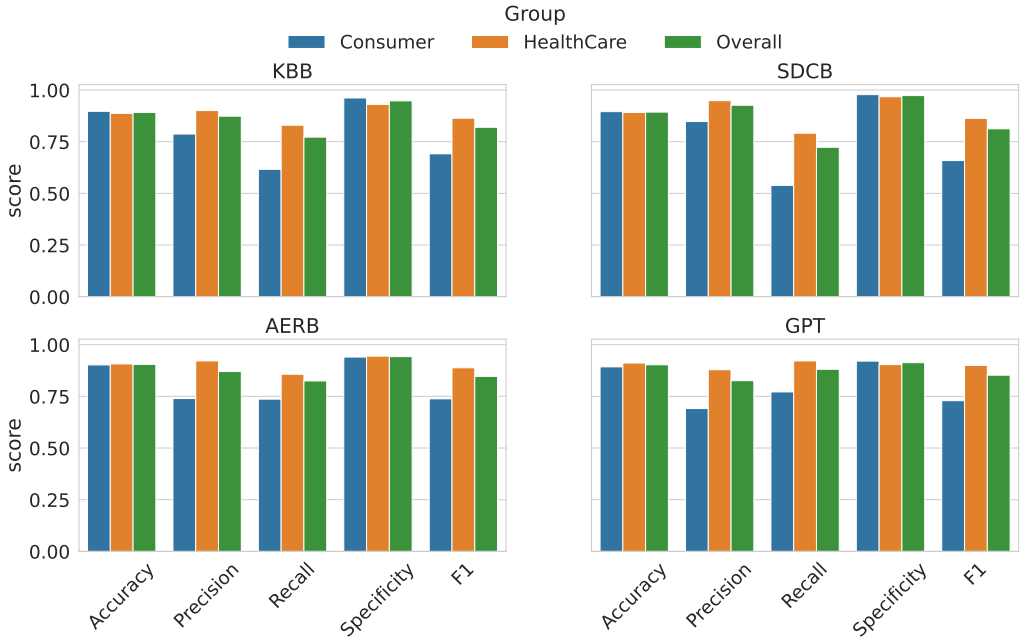


Figure 12: Model results on development data for reporter subgroups on partially corrected gold labels.

Figure 13: t-SNE projection of serious terms in Swedish ADRs according to IG attributions for four triage models. All terms are encoded with the same SentenceBERT model and each term is plotted individually as a point for each model. Manually assigned English cluster labels are added for the centroid of each cluster. The size of the points represents the spread of the cluster it belongs to specific to the explanations of a particular model. Terms occurring in the top lists of multiple models are represented as gradually more transparent points. Outliers are smallest and the most transparent.

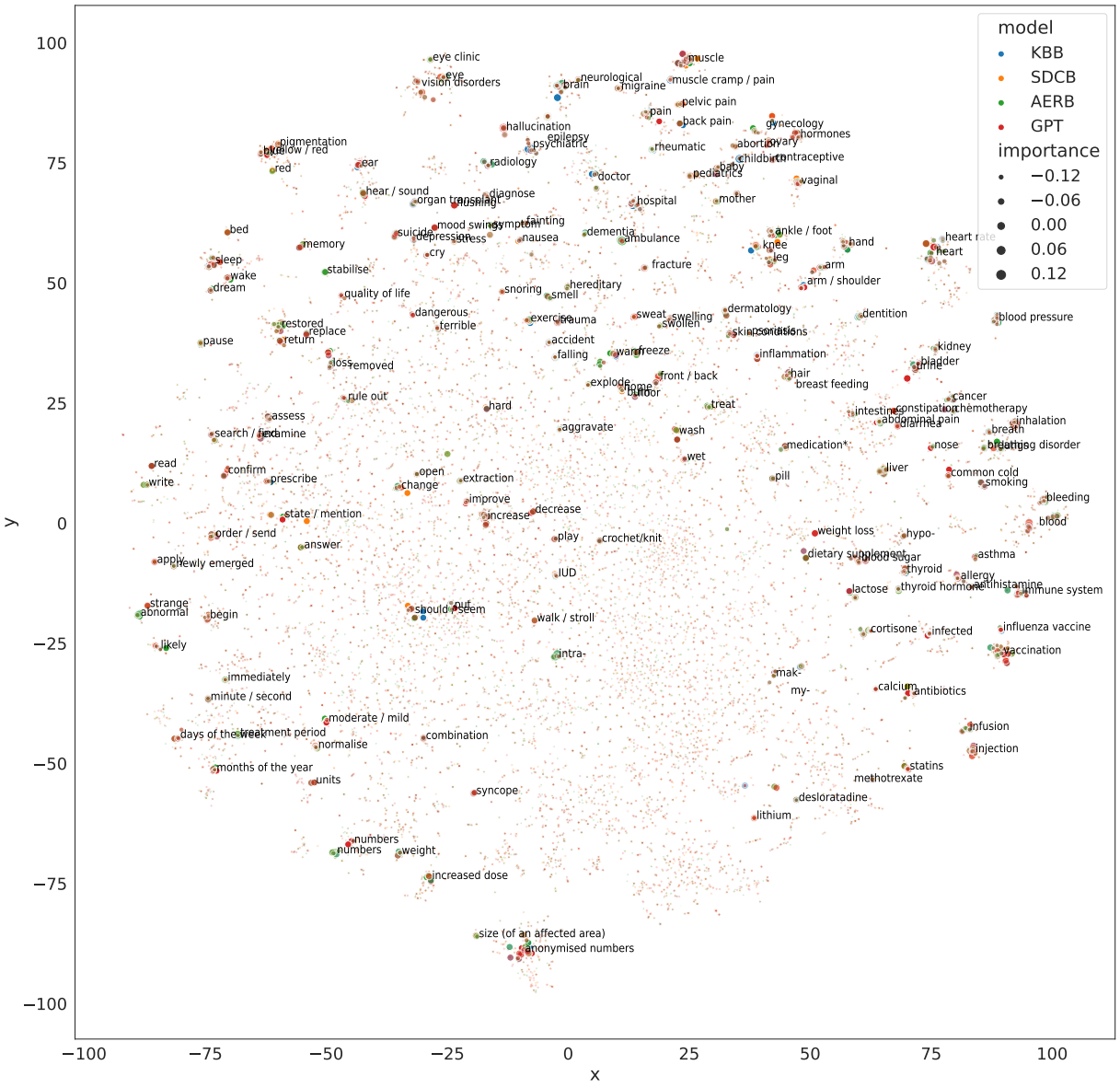


Figure 14: t-SNE projection of serious terms in Swedish ADRs according to EG attributions for four triage models. All terms are encoded with the same SentenceBERT model and each term is plotted individually as a point for each model. Manually assigned English cluster labels are added for the centroid of each cluster. The size of the points represents the spread of the cluster it belongs to specific to the explanations of a particular model. Terms occurring in the top lists of multiple models are represented as gradually more transparent points. Outliers are smallest and the most transparent.