Evaluating Explainability with LLM for Classification of Patient Safety Event Reports

Zhifei Dou

Supervisor: Eldan Cohen

Department of Mechanical and Industrial Engineering

University of Toronto

April 14, 2025

Contents

1 Abstract **3**

2 Introduction **4**

3 Related Work **4**

4 Data and Data Processing **5**
4.1 PSE Report Dataset Collection 5
4.2 Data Preprocessing 5

5 Method **5**
5.1 Prompt-based Generative Model Method 5
5.1.1 Model Selection for Generative Model Method 5
5.1.2 Classification Methods in Generative Model Method 6
5.1.3 Models' Explanations in Generative Model Method 8
5.1.4 Evaluation of Explanations in Generative Model Method 8

6 Experiment Result and Discussion **11**
6.1 Experiment Result for Generative Model Method 11
6.1.1 Classification Result 11
6.1.2 Evaluation Result 12

7 Conclusion and Future Work **13**

1 Abstract

Patient Safety Event (PSE) reports are essential for healthcare quality improvement; however, their classification presents challenges in consistency and correctness. While machine learning (ML) offers various solutions, traditional methods usually lack effective explainability, resulting in limited clinical utility and potentially biased insights. In this project, we explore the use of generative Large Language Models (LLMs), specifically small-to-medium scale open-sourced models (Qwen and its variants), for classifying PSE reports and generating Natural Language Explanations (NLEs). Addressing requirements of cost and confidentiality, modes are deployed locally. Prompting techniques: zero-shot, Retrieval Augmented Generation (RAG), and self-consistency are employed. Explainability is evaluated utilizing faithfulness tests (counterfactual editing) and factual correctness assessments (BioBERT entailment, FActScore). Results indicate that the tested LLMs have achieved classification performance comparable to the baseline ML models. Instruction-tuned and reinforcement-learning-trained models have demonstrated better faithfulness in their explanations compared to distilled models. On the other hand, evaluations also suggest potential issues with the factual correctness of the generated NLEs, emphasizing the demand for robust validation. This work has underscored the potential of LLMs in PSE report analysis while suggesting the critical importance of evaluating the reliability of explanations generated by these models.

2 Introduction

Patient Safety Report (PSE) reports are documents that record unexpected events, errors, accidents or situations that affect or potentially affect a patient's care in hospital [1]. Such reports are a way healthcare organizations can capture incidents that can be reviewed and improved in the future. The publication of the study 'To Err Is Human' has demonstrated to the public the essentialness of preventing patient injuries; thus, also identifying the essentialness of correct classification for these reports[2].

Currently, there have been several works on correctly classifying the PSE reports with machine learning models. For example, previous works done by Chen et al. have implemented a comprehensive evaluation for machine learning methods (SVM, K-nearest Neighbours) with static embeddings (Word2Vec, GloVe, etc.) and also context embeddings (BERT embedding, RoBERTa embedding)[3].

However, even though Chen et al. have adopted an explainability tool - LIME after classification to explain how the prediction is made by the machine learning model, such an explanation can only highlight the 'critical' words in the input PSE report rather than Natural Language Explanation (NLE). The study from Yanagawa et al. has shown that showing traditional explanation results such as highlighting critical words or saliency maps has a limited benefit on clinical performance[4]. The clinical usefulness of such an explainability tool is thus doubtful. The study by Ghassemi et al. points out that traditional explainability tools sometimes even distract doctors from the true essential sections; thus, producing biased judgement[5]. Also, the evaluation method proposed by Chen et al. has only focused on the factual correctness of the explainability tools, leaving question marks about the faithfulness of the experimented models. In other words, we don't know if the explanation of the model reflects its true reasoning path [3].

As transformer-decoder auto-regressive Large Language Models (LLM), such as QwQ from Alibaba AI, have demonstrated a decent ability in the medical domain and perform general text classification tasks, adopting such models can potentially have better classification and explanations than the machine-learning models[6, 7]. Hence, in this project, the team has comprehensively adopted GPT-like generative LLMs to classify PSE reports. Considering the confidential issue of using API and the expensive cost of deploying full-size LLMs, small to middle-scale open-sourced LLMs are mainly deployed locally in this project.

Meanwhile, NLE is generated along with the predicted label from the LLM to reveal the rationales behind the models' decisions. By providing NLE as an explanation, a more human-like and complete rationale is provided and potentially augments the clinician's knowledge; thus solving the issue of traditional explainability tools[8]. A study by Kayser et al. has also confirmed textual explanations are more useful than purely visual ones for clinicians[9]. On the other hand, studies have shown that NLE is also facing challenges and potentially generates hallucinations[10]. Thus, comprehensive evaluation metrics, such as counterfactual tests and FActScore, are implemented along with human evaluations to validate the explanation's quality such as faithfulness and factual correctness.

As a result, all generative LLMs tested have demonstrated decent performance in classification which is compatible with the baseline models. For explanation, instruction models and reinforcement-learning trained reasoning models demonstrate more faithful explanations than the distilled models.

The code of this project can be found in:
`https://github.com/ZhifeiDou/Explanability-in-PSE-report-Classification.git`

3 Related Work

As introduced, prior research by Chen et al. has explored the classification of PSE reports with various machine learning models, including Support Vector Machines (SVM) and K-nearest Neighbors, combined with different text embedding techniques, such as static Word2Vec, GloVe, and contextual BERT, RoBERTa embeddings. The classification performance achieved in their study serves as a baseline for this project. [3].

Furthermore, methodologies for evaluating the quality of NLE generated by models have been proposed in many other studies. Specifically, the work proposed by Atasnova et al. and Valentino et al. introduced methods to evaluate NLEs [8, 11]. These works inform the evaluation strategies employed in this project for assessing the explanations generated by the LLMs.

4 Data and Data Processing

4.1 PSE Report Dataset Collection

The PSE report dataset utilized in this project was obtained from 'Labor and Delivery' and 'Mother-baby' units of an academic hospitality organization located in the southeast of the United States. The dataset consists of 861 PSE reports which were issued from January 1st, 2019 to December 31st, 2020, along with the class labels. All PSE reports are anonymized according to the privacy regulation requirement.

4.2 Data Preprocessing

Then, the 'Report Description' and 'Event Type' are extracted as the free-text feature and target. To avoid sampling bias, only major classes are kept. The kept classes and the number of data pieces they contain are demonstrated in the following table:

Table 1: Processed PSE Dataset Classes and Frequencies

Class Name	**Frequency**
Care coordination / communication	186
Laboratory test	122
Medication related	89
Omission / errors in assessment, diagnosis, monitoring	67
Maternal	58
Equipment / devices	56
Supplies	49

The dataset has been split into training, validation, and testing subsets with each having portions of 60%, 20%, and 20%.

5 Method

5.1 Prompt-based Generative Model Method

Unlike the previous work, this project has prompted generative LLMs with an auto-regressive generation head as the last layer. In this section, we have discussed the rationales for the selection of generative LLMs, classification methods, and evaluation methods. All technical details are provided in the Appendix.

5.1.1 Model Selection for Generative Model Method

Currently, the latest full-size generative LLMs, such as ChatGPT-4o and Deepseek-V3, are usually deployed on cloud service. Meanwhile, PSE reports are usually confidential to protect the personal information of patients and healthcare staff [12]. Access to such models through cloud services has the risk of information leakage[13]. On the other hand, deploying such full-size generative LLMs locally is not always affordable for all healthcare organizations[14]. Also, recent studies from Qiu et al. have shown even small or middle reasoning models performed decently in the medical domain's tasks[7]. Hence, the team proposed experimenting with the performance of open-sourced small-scale models such as LLaMA-8B from Meta and Qwen-14B from Alibaba [15][16]. Also, the model distillation technique allows small-scale LLMs to learn reasoning ability from full-size LLMs. Deepseek has released distilled Qwen models from their first reasoning model

Deepseek-R1, which potentially produces better explanations for the PSE report classification task[17]. Similarly, Alibaba has released its first reasoning Qwen: Qwen with Question (QwQ) highlighting outstanding performance in reasoning benchmarks[6]. Hence, the team has chosen models demonstrated in the following table 2.

Table 2: LLMs with Generative Head Chosen in this Project

Model Name	**Number of Parameters (Billion)**
Qwen - 2.5	14 and 32
QwQ	32
Deepseek Distilled Qwen	14 and 32

5.1.2 Classification Methods in Generative Model Method

All the following tasks are implemented on a single RTX4090 24G, 4-bit quantization, all LLMs and their tokenizers are downloaded from Huggingface. The seed is fixed to 42 for reproducibility. Due to the lack of a human-labelled golden explanation, instruction-based learning is the most feasible learning strategy for this project. The team has prompted the models listed in table 2 in zero-shot, few-shot, and Retrieval Augmented Generation(RAG) format along with techniques including explanation-based prompting, self-consistency, Chain-of-Thought(CoT) and persona embedding to increase the model's performance in a classification task. To make the result comparable with the other project, the testing dataset is used for the classification task which contains 124 PSE reports. All Qwen instruction series models are applied to chat templates for proper outputs. The input and output can be visualized in figure 1

Figure 1: Generative Model I/O for Classification Task

Zero-shot Prompting Zero-shot prompting is the technique where generative LLMs are given an instruction prompt to perform the task without examples. The zero-shot prompt (base prompt) is demonstrated in the main body table 3. For zero-shot and RAG in the following, we adopted Top-P as 0.9 and temperature as 0.7 for all models due to the common recommendations from models using guidance[6, 16].

Retrieval-Augmented Generation Retrieval-Augmented Generation proposed by Lewis et al. is a framework that combines the generative LLMs with a retrieval mechanism[18]. In this project, we use a separate BioBERT model's embedding as the representation due to its embedded biomedical knowledge within the embedding[19]. The PSE report contents of the training dataset are mapped into BioBERT embedding. We use Faiss from the Meta AI team as the similarity search tool to retrieve the closest content from the training dataset to the given one. Then, the following prompt demonstrated in table 3 is input into the generative LLM with the retrieved content.

Self-consistency Self-consistency is a decoding strategy designed to improve the reasoning abilities of generative LLMs. One of the most effective self-consistency methods proposed by Wang et al. is making generative LLMs generate multiple distinct reasoning paths for a given task or problem and perform majority voting to determine the final decision [20]. Following the original work, we have also performed hyperparameter tuning on key factors of temperature, Top K, and the number of reasoning paths sampled for each model and elect the best-performed hyperparameter settings. The results show that the best setting is temperature = 0.7, Top K = 40, and 10 as the sample number. The prompt in table 3 is also used. The majority voting process can be demonstrated as the following functions:

$$\text{Given sampled\_classes} = \{\, c_1, c_2, \ldots, c_n\} \quad \text{and} \quad \text{sampled\_explanations} = \{\, e_1, e_2, \ldots, e_n\},$$

$$\text{let } \text{freq}(c) = \sum_{i=1}^{n} \mathbf{1}\big(c_i = c\big).$$

$$\text{Then, final\_class} = \arg\max_{c} \text{freq}(c),$$

$$\text{and we define final\_explanation} = e_j \quad \text{where } j = \min\Big\{\, i \mid c_i = \text{final\_class}\Big\}.$$

Table 3: LLMs Prompting for Classification and Explanation

Prmpt: Base prompt adopted (definition for each class omitted)
You are a medical classification assistant. Your task is to read the text below and decide which of the following categories best applies: if RAG: Below are some relevant records from our training dataset (content + label): {retrieved content} Categories (with brief definitions): Care coordination / communication: Incidents... Laboratory test: Definition: Laboratory test-related incidents ... Medication related: Medication-related incidents ... Omission / errors in assessment or diagnosis or monitoring: These incidents ... Maternal: Maternal safety incidents ... Equipment / devices: This category covers ... Supplies: Supplies-related incidents ... **Instructions**: 1. Read the text carefully and identify the key points or keywords that hint at one of the above categories. 2. Decide on the single most relevant category. 3. Provide a short explanation of how you arrived at this decision, referencing specific keywords or ideas from the text. 4. Think step by step like a professional doctor 5. Return the result in **exactly** the following Markdown format: **Class:** Your single best class here **Explanation:** Your explanation here Text: {PSE report content}

5.1.3 Models' Explanations in Generative Model Method

Through the prompting shown in table 3, the model is also introduced to articulate a step-by-step reasoning path in parallel with predicting the classification labels. This procedure yields an NLE alongside the prediction, producing a human-interpretable text rationale that demonstrates the model's decision-making process.

5.1.4 Evaluation of Explanations in Generative Model Method

Faithfulness For generative models, faithfulness measures if the explanation truly reflects the model's actual decision process [21]. To measure faithfulness, the team has set up a counterfactual editor and edited the input to see how many predictions changed the tested model, following the work proposed by Atanasova et al [8]. According to the proposed method, the team has adopted T5 as the base model for the counterfactual editor. The T5 model is fine-tuned with the following three losses: 1. Mask-filling loss. 2. Imitation loss. 3. Adversarial loss [8]. With this fine-tuned counterfactual editor, we produce and feed counterfactual PSE report content into the tested models. If the tested model's explanation is faithful, the model's prediction remains unchanged and vice versa [8].

Like the author, the counterfactual editor model we adopt is a T5-based generative model proposed by Google Brain and fine-tuned with the following loss[22].

First is the filling loss, where we randomly mask the PSE reports' word in content and let the T5 model predict the masked word. The cross-entropy loss of predicting this masked word is the filling loss demonstrated as the following equation.

$$\mathcal{L}_{\text{filling}} = -\sum_{t=1}^{|W|} \log\Big(p_\theta\big(w_t \mid w_{<t}, \text{Masked}(x_i)\big)\Big)$$

where $W = (w_1, w_2, \ldots, w_{|W|})$ is the sequence of ground-truth tokens that were masked, $p_\theta(\cdot)$ is the T5 probability distribution over vocabulary tokens, $\text{Masked}(x_i)$ denotes the original input x_i with a contiguous substring replaced by a placeholder, and we optionally condition on the current or target label y_i^C as an input prefix.

Then, we replace the mask with the predicted word by the T5 model and feed this processed PSE report content into the model we would like to test the faithfulness on, calling it the teacher model. We then take the explanation generated by the teacher model and let the T5 model mimic the teacher model's explanation. In other words, we perform a teacher forcing to the T5 model with the teacher model's explanation. The cross-entropy loss during this teacher forcing is called imitation loss, represented by the following function.

$$\mathcal{L}_{\text{imitation}} = -\sum_{t=1}^{|E^{(T)}|} \log\Big(p_\theta\big(E_t^{(T)} \mid E_{<t}^{(T)}, X\big)\Big)$$

where $E^{(T)} = \big(E_1^{(T)}, E_2^{(T)}, \ldots, E_{|E^{(T)}|}^{(T)}\big)$ is the teacher-provided explanation (token sequence), $p_\theta(\cdot)$ is the T5 probability distribution over vocabulary tokens, and X denotes the model's input which is processed PSE report content.

Finally, we take the logits of the mask-filling task and the explanation mimic task to calculate the difference between them, the negation of the difference is the adversarial loss which helps us to push these two logits as far away as possible. Hence, once we finish training, the text generated by the T5 model will be words that make the original content as far away from the explanation as possible. The adversarial loss $\mathcal{L}_{\text{adv}}$ can be demonstrated as the following function:

$$\text{fill\_mean}_b = \frac{1}{|\mathbf{F}_b|}\sum_{j=1}^{|\mathbf{F}_b|}\text{logits\_infill}_{b,j} \quad , \quad \text{expl\_mean}_b = \frac{1}{|\mathbf{E}_b|}\sum_{j=1}^{|\mathbf{E}_b|}\text{logits\_expl}_{b,j}$$

$$\text{diff}_b \;=\; \Big|\,\text{fill\_mean}_b \;-\; \text{expl\_mean}_b\Big|, \quad \overline{\text{diff}} \;=\; \frac{1}{B}\sum_{b=1}^{B}\text{diff}_b$$

$$\mathcal{L}_{\text{adv}} \;=\; \Big(-\,\overline{\text{diff}}\Big)$$

where $\text{logits\_infill}_{b,j}$ and $\text{logits\_expl}_{b,j}$ denote the respective logit values from (i) the editor's fill-generation pass and (ii) the editor's teacher-explanation pass, $|\mathbf{F}_b|$ and $|\mathbf{E}_b|$ are the total logit counts in each pass, B is the batch size, and λ_{adv} is the adversarial weight hyperparameter.

Then we put it all together:

$$\mathcal{L}_{\text{editor}} \;=\; \lambda_f\,\mathcal{L}_{\text{filling}} \;+\; \lambda_i\,\mathcal{L}_{\text{imitation}} \;+\; \lambda_a\,\mathcal{L}_{\text{adv}},$$

where $\mathcal{L}_{\text{filling}}$ is the cross-entropy filling loss, $\mathcal{L}_{\text{imitation}}$ is the teacher-forcing imitation loss, $\mathcal{L}_{\text{adv}} = -\overline{\text{diff}}$ is the adversarial term, and $\lambda_f, \lambda_i, \lambda_a$ are their respective weighting factors as hyperparameters.

Then the team has adopted this fine-tuned T5 model to edit the original PSE report content to produce counterfactual PSE report content for each model tested. The counterfactual content is used for each model to inference and produce predictions for counterfactual content. The number of differences between new predictions and original predictions directly correlated with the extent of the tested model's faithfulness. The procedure of fine-tuning the counterfactual is visualized in figure 2.

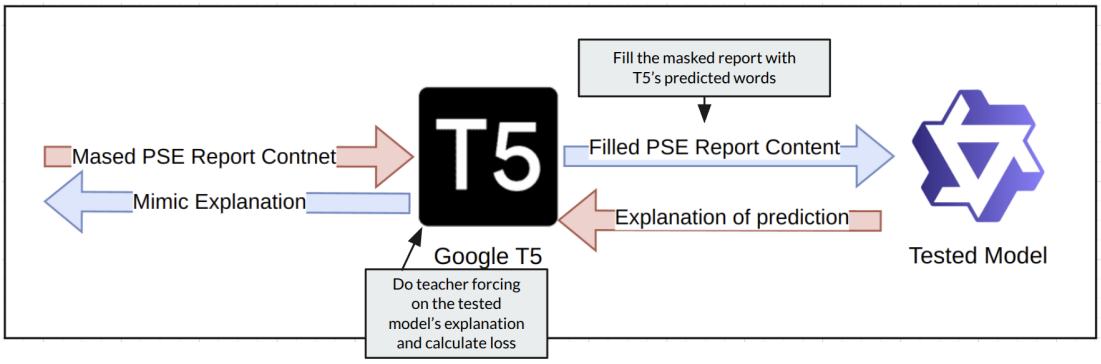

Figure 2: Fine-tuning Procedure Map for Counterfactual Editor

Factual Correctness Then we also need to evaluate the factual correctness of the explanation, which we propose to define as whether the explanation is true to the facts of the PSE report as well as the general medical knowledge. In this project, we have used three methods to evaluate this metric which are: the **BioBERT entailment method**, the **FActScore method**[23, 24].

The BioBERT entailment method, proposed by Jeong et al., is under the suggestion that: If an explanation is factually correct, it should have an entailment relationship with the PSE report content[23]. Work proposed by Valentino et al. also validated if an explanation is factually correct, it should be entailed by the input content[11]. Hence, following the method of Jeong et al., the team has fine-tuned a BioBERT model with the well-known e-SNLI entailment dataset proposed by Camburu et al[25] to test the factual correctness.

The BioBERT model is fine-tuned with the following cross-entropy loss function.

$$\mathcal{L}_{\text{CE}}(\theta) = -\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbf{1}[y_i = c]\,\log\Big(p_\theta\big(c \mid x_i\big)\Big),$$

where x_i is the input text, which is premise and hypothesis from e-SNLI dataset, y_i is the true label which is one of the entailment, neutral, and contradiction, in e-SNLI. $p_\theta(c \mid x_i)$ is the predicted probability of class c under the model parameters θ.

After fine-tuning, the original PSE report content and the explanation generated by the tested model are formed as the premise and hypothesis input to the BioBERT model. If the BioBERT model's prediction is settled in neutral or entailment, it implies the explanation is factually correct, if the prediction is a contradiction, it implies the explanation potentially contains hallucination information. The procedure for using the fine-tuned BioBERT is represented in figure 3.

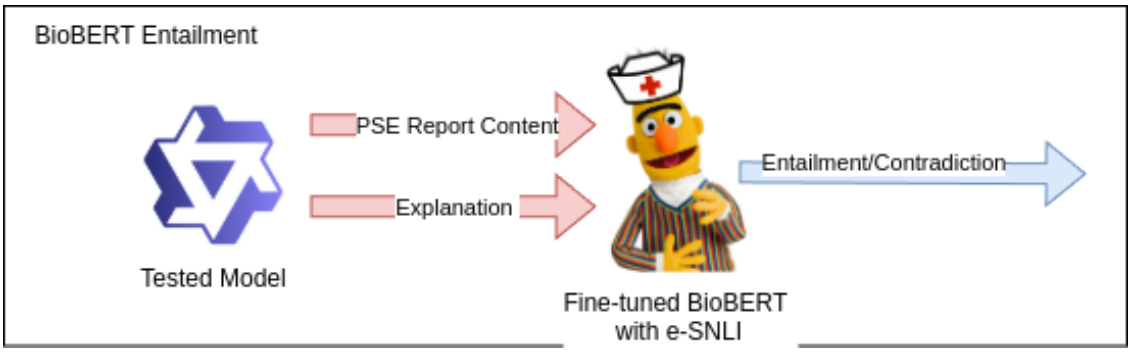

Figure 3: Fine-tuned BioBERT Entailment Procedure

Meanwhile, the team also adopted FActScore proposed by Min et al., which breaks the explanation into atomic facts and fact checks, to further validate the factual correctness of the explanation [24]. Due to the hardware constraint, we adopted Qwen14B as the assist model which is in charge of forming atomic facts and validating factual correctness. Following the original method, we have locally stored and indexed Wikipedia via pyserini. For each atomic fact, we retrieve the closest wiki content by BM25 as suggested in the original paper. Then the atomic fact and wiki content are input into the validator model, the model makes its judgement if the atomic fact is true according to the content. After all atomic facts are judged, the percentage of true predictions within all predictions would be the representation of model explanations' factual correctness. The FActScore process can be visualized with the following figure 4.

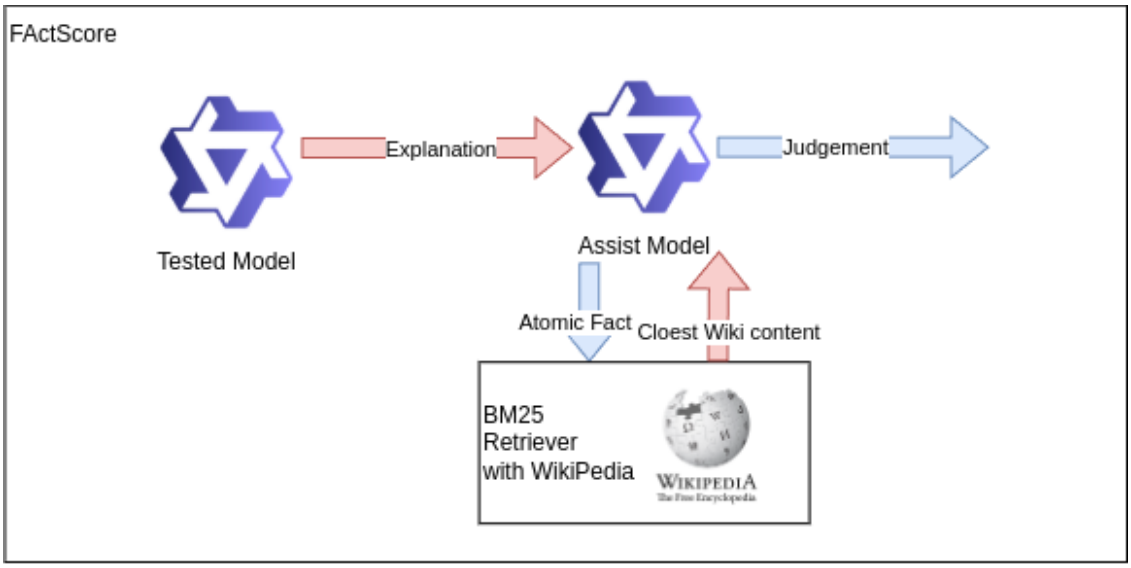

Figure 4: FActScore Evaluation Procedure

6 Experiment Result and Discussion

6.1 Experiment Result for Generative Model Method

6.1.1 Classification Result

Table 4: Classification Metrics for Each Model (EP: Empty Prediction, SC: self-consistency, ZS: Zeroshot)

Model	Accuracy	Precision	Recall	F1 Score	EP	Generation Time
Qwen14B_ZS	0.6935	0.5884	0.5685	0.5719	0	5'22"
Qwen32B_ZS	0.6452	0.6144	0.6040	0.5867	0	9'45"
QwQ_ZS	0.6935	0.6895	0.6667	0.6635	0	47'30"
Deepseek_Distilled_Qwen14B_ZS	0.5081	0.5519	0.4823	0.4795	1	5'58"
Deepseek_Distilled_Qwen32B_ZS	**0.7097**	0.6040	0.5886	0.5844	0	10'35"
Qwen14B_SC	0.6774	**0.6879**	0.6350	0.6401	0	54'10"
Qwen32B_SC	0.6452	0.6016	0.6017	0.5842	0	108'43"
QwQ_SC	0.7016	0.6717	**0.6704**	**0.6610**	0	501'6"
Deepseek_Distilled_Qwen14B_SC	0.4274	0.6171	0.3575	0.3523	9	288'28"
Deepseek_Distilled_Qwen32B_SC	**0.7097**	0.6664	0.6665	0.6587	0	260'10"
Qwen14B_RAG	0.6774	0.6339	0.5942	0.6040	0	5'23"
Qwen32B_RAG	0.6935	0.6226	0.6173	0.6141	0	10'28"
QwQ_RAG	**0.7097**	0.5824	0.5931	0.5856	0	47'37"
Deepseek_Distilled_Qwen14B_RAG	0.6935	0.5781	0.5532	0.5527	0	6'30"
Deepseek_Distilled_Qwen32B_RAG	0.7016	0.5590	0.5624	0.5564	0	32'45"

The classification result is demonstrated in the table 4. All tested models have demonstrated similar performance except for distilled Qwen14B which generally has the worst performance with empty predictions.

Reasoning models distilled Qwen32B and QwQ, have generally demonstrated better performance than non-reasoning Qwen32B, suggesting the effectiveness of reasoning ability given by distillation and reinforcement learning (RL) processes. On the other hand, QwQ is suffering from a long thinking time, compared with the other 32B scaled models due to its unique math and coding fine-tuning RL process[6].

Comparing the instruction-learning techniques, larger scale models with self-consistency have achieved top classification performance in all accuracy, precision, recall, and f1-score metrics. It identifies the effectiveness of self-consistency in boosting the generative LLMs' classifying ability. However, it also takes longer generation time which is approximately N (sample number we take vote from) times longer than zero-shot generation. Also, self-consistency decreases the performance of smaller models which fail to generate the correct output format. We can observe this from Deepseek distilled Qwen14B which generates a single unformatted output by itself, but generated 9 unformatted outputs when implementing self-consistency which aligns with the finding from Wang et al [20]. Meanwhile, RAG generally improved performance over the zero-shot baseline for most models while maintaining the similar generation times, demonstrating its effectiveness in leveraging relevant examples from the training data.

Comparing to the baseline results reported by Chen et al., the LLMs used in this study have shown competitive performance. The baseline's top performance classifier using static text representation (MLR with TF-IDF) has achieved an accuracy of 66.7% and the F1-score of 0.631 [3]. Several models in this study, including Qwen14B, QwQ (with zero-shot and self-consistency), Deepseek Distilled Qwen32B (also with zero-shot and self-consistency), and various RAG configuration models, outperformed this static baseline. However, the baseline study also reported a higher benchmark using contextual embeddings (SVM with RoBERTa-base model), which has achieved an accuracy of 75.4% and an F1-score of 0.753[3]. The LLMs tested in this project, while also demonstrating relatively strong performance and surpassing the static baseline, did not reach this higher contextual baseline. This suggests that even though generative LLMs

offer advantages in terms of generating NLEs, further optimization or domain-specific model choices might be needed to match the peak classification accuracy achieved by specialized classifier models.

6.1.2 Evaluation Result

Table 5: Faithfulness Result for Each Model - Generative Model (EP: Empty Prediction, SC: self-consistency, ZS: Zeroshot)

Model	Changed Prediction Number
Qwen14B_ZS	11
Qwen32B_ZS	12
QwQ_ZS	12
Deepseek_Distilled_Qwen14B_ZS	35
Deepseek_Distilled_Qwen32B_ZS	18
Qwen14B_SC	11
Qwen32B_SC	15
QwQ_SC	5
Deepseek_Distilled_Qwen14B_SC	53
Deepseek_Distilled_Qwen32B_SC	23
Qwen14B_RAG	10
Qwen32B_RAG	14
QwQ_RAG	18
Deepseek_Distilled_Qwen14B_RAG	22
Deepseek_Distilled_Qwen32B_RAG	25

Faithfulness From the result table 5, we can observe that Deepseek distilled models generally have more prediction changes with the counterfactual dataset which potentially implies the distilled models are less faithful than the others. The team suspect this is caused by the distillation process where the student model has been forced to think in the teacher model's reasoning path and then make a prediction, but the taught reasoning path is likely memorized by the student models rather than producing the student model's own reasoning path. This finding aligns with the finding from Wang et al. where the simple supervised fine-tuning distillation, which is the distillation technique adopted for distilled Qwens, would produce an unfaithful student model [26, 17].

In contrast, the other reasoning model trained with RL, QwQ, is generally equally faithful with instruction fine-tuned Qwens according to their relatively low changed prediction numbers, demonstrating their solid faithful reasoning path to predictions.

Also, self-consistency can generally impact the model's faithfulness. As shown by the result of distilled Qwen14B, self-consistency can highlight less faithful models by amplifying the testing result. This observation is reasonable because by taking majority voting, multiple generations from an unfaithful model would produce a more chaotic explanation than a single one.

Table 6: Factual Correctness - BioBERT Entailment Result for Each Model - Generative Model (EP: Empty Prediction, SC: self-consistency, ZS: Zeroshot)

Model Name	**Contradiction Count**	**FActScore**
Qwen14B_ZS	85	33.88%
Qwen32B_ZS	89	33.53%
QwQ_ZS	96	31.97%
Deepseek_Distilled_Qwen14B_ZS	97	32.35%
Deepseek_Distilled_Qwen32B_ZS	97	33.53%
Qwen14B_SC	81	33.96%
Qwen32B_SC	88	34.90%
QwQ_SC	101	33.66%
Deepseek_Distilled_Qwen14B_SC	93	31.66%
Deepseek_Distilled_Qwen32B_SC	96	33.69%
Qwen14B_RAG	84	32.26%
Qwen32B_RAG	82	31.13%
QwQ_RAG	91	31.60%
Deepseek_Distilled_Qwen14B_RAG	91	32.40%
Deepseek_Distilled_Qwen32B_RAG	87	32.33%

Factual Correctness - BioBERT Entailment As the table 6 demonstrates, the contradiction count is the number of contradiction predictions from fine-tuned BioBERT to the generative LLM's explanation. The total prediction number is 124. We can observe that all models tested have demonstrated a high contradiction count, indicating the significant existence of false information within the explanation.

This observation can be further verified by the FActScore's result. Besides the contradiction count is the FActScore for each model, none of the scores have exceeded 34% identifying one-third of the atomic facts are not supported by the facts stored in Wikipedia. As a baseline, ChatGPT has obtained 58% FActScore in the work proposed by Min et al[24]. It further suggests all models tested potentially failed to provide true information resulting in poor factual correctness.

These results are potentially caused by the lack of domain-specific knowledge for tested models. As Qwen series and its variants are trained on board, general text corpus, these datasets may lack deep and nuanced understanding of specific medical terminology, hospital procedures, or the particular context regarding PSE reports. Especially our PSE reports often contain a significant number of jargons, abbreviations, and descriptions of situations which require specialized knowledge to interpret and explain accurately.

But the team would also like to argue that this observation needs to be further validated with an ablation study on BioBERT Entailment and FActScore. The BioBERT entailment methods adopted the e-SNLI dataset in the original method rather than the medical domain inference dataset; thus, the fine-tuned BioBERT model potentially lacks medical domain-specific inference knowledge. For the FActScore, due to the hardware and confidential constraint, Qwen14B is used as the assistant generative LLM. Compared with the full-size ChatGPT adopted in the original FActScore paper, Qwen14B potentially lacks of power to perform proper judgement on the provided atomic facts and Wikipedia contents[24]. Thus, potentially leading to incorrect judgment.

7 Conclusion and Future Work

This project investigated the application of generative LLMs for classifying Patient Safety reports and generating accompanying natural language explanations. The study found that tested open-source LLMs, using instruction-learning settings(zero-shot, RAG, self-consistency), showed classification capabilities comparable to established ML baselines. A critical finding related to explainability: non-distilled models, especially

for those fine-tuned with reinforcement learning (QwQ), have produced explanations that more faithfully reflected their internal decision processes compared with models created via simple supervised fine-tuning distillation. Als,o assessments for factual correctness with BioBERT entailment and FActScore revealed major concerns, with high contradiction rates and low scores demonstrating potential inaccuracies or hallucinations with the explanation generated.

Such findings have highlighted a trade-off: while LLMs offer promising performance and more intuitive NLEs, ensuring the reliability and factual grounding of such explanations remains challenging without fine-tuning using the medical-domain-specified corpus. Hence, in the future, works should focus on further validating the factual correctness results, potentially employing larger and more domain-specific models for the entailment model or FActScore assistant. Also, manual evaluation of both the classification labels and the generated explanations by clinical domain experts is necessary to ensure the practical utility and trustworthiness of these models in real-world PSE contexts.

References

[1] S. Alder et al. What is incident reporting in healthcare? https://www.hipaajournal.com/what-is-incident-reporting-in-healthcare/, 2024. Accessed: April 13, 2025.

[2] L. Kohn et al. To err is human: Building a safer health system. https://pubmed.ncbi.nlm.nih.gov/25077248/, 2000. Accessed: April 13, 2025.

[3] H. Chen et al. A machine learning approach with human-ai collaboration for automated classification of patient safety event reports: Algorithm development and validation study. https://humanfactors.jmir.org/2024/1/e53378/, 2024. Accessed: April 13, 2025.

[4] M. Yanagawa et al. Seeing is not always believing: Discrepancies in saliency maps. https://pmc.ncbi.nlm.nih.gov/articles/PMC10831517/, 2023. Accessed: April 13, 2025.

[5] M. Ghassemi et al. The false hope of current approaches to explainable artificial intelligence in health care. https://www.thelancet.com/journals/landig/article/PIIS2589-7500(21)00208-9/fulltext, 2021. Accessed: April 13, 2025.

[6] Qwen Team. Qwq-32b: Embracing the power of reinforcement learning. https://qwenlm.github.io/blog/qwq-32b/, 2025. Accessed: April 13, 2025.

[7] P. Qiu et al. Exploring the limits of transfer learning with a unified text-to-text transformer. https://arxiv.org/abs/2503.04691, 2025. Accessed: April 13, 2025.

[8] P. Atanasova et al. Faithfulness tests for natural language explanations. https://arxiv.org/abs/2305.18029, 2023. Accessed: April 13, 2025.

[9] M. Kayser et al. Fool me once? contrasting textual and visual explanations in a clinical decision-support setting. https://arxiv.org/abs/2410.12284, 2024. Accessed: April 13, 2025.

[10] A. Madsen et al. Are self-explanations from large language models faithful? https://arxiv.org/abs/2401.07927, 2024. Accessed: April 13, 2025.

[11] M. Valentino et al. Do natural language explanations represent valid logical arguments? verifying entailment in explainable nli gold standards. https://arxiv.org/abs/2105.01974, 2021. Accessed: April 13, 2025.

[12] L. Ginsburg et al. Development of a measure of patient safety event learning responses. https://pmc.ncbi.nlm.nih.gov/articles/PMC2796318, 2009. Accessed: April 13, 2025.

[13] S. Kim et al. Propile: Probing privacy leakage in large language models. https://arxiv.org/abs/2307.01881, 2023. Accessed: April 13, 2025.

[14] N. Alkhaldi et al. Assessing the cost of implementing ai in healthcare. https://itrexgroup.com/blog/assessing-the-costs-of-implementing-ai-in-healthcare/, 2024. Accessed: April 13, 2025.

[15] A. Grattafiori et al. The llama 3 herd of models. https://arxiv.org/abs/2407.21783, 2024. Accessed: April 13, 2025.

[16] A. Yang et al. Qwen2.5 technical report. https://arxiv.org/abs/2412.15115, 2025. Accessed: April 13, 2025.

[17] Deepseek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. https://arxiv.org/abs/2402.12749, 2025. Accessed: April 13, 2025.

[18] P. Lewis et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. https://arxiv.org/abs/2005.11401, 2020. Accessed: April 13, 2025.

[19] J. Lee et al. Biobert: a pre-trained biomedical language representation model for biomedical text mining. https://arxiv.org/abs/1901.08746, 2019. Accessed: April 13, 2025.

[20] X. Wang et al. Self-consistency improves chain of thought reasoning in language models. https://arxiv.org/abs/2203.11171, 2023. Accessed: April 13, 2025.

[21] C. Agarwal et al. Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models. https://arxiv.org/abs/2402.04614, 2024. Accessed: April 13, 2025.

[22] C. Raffel et al. Exploring the limits of transfer learning with a unified text-to-text transformer. https://arxiv.org/abs/1910.10683, 2019. Accessed: April 13, 2025.

[23] M. Jeong et al. Olaph: Improving factuality in biomedical long-form question answering. https://arxiv.org/abs/2305.18029, 2025. Accessed: April 13, 2025.

[24] S. Min et al. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. https://arxiv.org/abs/2305.14251, 2023. Accessed: April 13, 2025.

[25] O. Camburu et al. e-snli: Natural language inference with natural language explanations. https://arxiv.org/abs/1812.01193, 2018. Accessed: April 13, 2025.

[26] P. Wang et al. Scott: Self-consistent chain-of-thought distillation. https://arxiv.org/abs/2305.01879, 2023. Accessed: April 13, 2025.