

Gooseek@SMM4H-HeaRD 2025: Detection of Personal Adverse Reactions to Shingles Vaccines in Reddit Posts using Large Language Models

Shichao Feng^{1†}, Donger Chen^{1†}, Yuan Li^{1*}, Bailu Zhang^{1*}

¹Department of Computer Science and Engineering
University of North Texas

3940 N Elm St, Denton, TX, 76207, USA

FengFeng@my.unt.edu, DongerChen@my.unt.edu, Yuan.Li2@unt.edu, BailuZhang@my.unt.edu

Abstract

Traditionally, pharmacovigilance (PV) relies on formal reporting systems or official databases, which often underreport patient experiences. Social media, particularly Reddit, offer valuable resources for real-time detection and tracking of personal adverse reactions, especially in the context of rapidly spreading epidemic diseases. However, extracting meaningful insights from Reddit posts remains challenging due to their inherent noise and unstructured nature. In this study, we leverage the Llama 3.1-8B-Instruct large language model (LLM), enhanced through supervised fine-tuning (SFT) and integrated with few-shot or zero-shot learning strategies using chain-of-thought (CoT) prompts, to automatically identify personal adverse reactions to shingles vaccines from Reddit discussions in the context of Task 6 of the SMM4H-HeaRD 2025 shared task. Experimental evaluations demonstrate the model’s capability to accurately and efficiently detect mentions of adverse reactions in Reddit posts, achieving F1-score of 96.3% on the holdout dataset. Additionally, an ablation study reveals that incorporating CoT prompts significantly improves overall performance, whereas few-shot learning somewhat introduces bias. The code is publicly available at https://github.com/ShichaoFeng92/SMM4H_2025_Task6.git

Introduction

Social media has become a popular platform for sharing personal experiences related to adverse reactions following medication intake or vaccine administration (Melton et al. 2021). Compared to conventional public health surveillance methods, social media offers real-time responsiveness, making it well-suited for monitoring emerging public health trends, particularly in the post-COVID-19 era (Kwon and Park 2023). However, due to the informal and flexible nature of user expression on these platforms, social media content often contains noise, such as slang, typos, sarcasm, and ellipses. Given the massive volume of daily posts, extracting relevant mentions of adverse reactions to shingles vaccines presents a significant challenge.

Pretrained language models (PLMs) like RoBERTa have been widely adopted for adverse reaction detection tasks due to their efficiency and performance in SFT settings (Barbieri et al. 2020; Antypas et al. 2023). PLMs are typically fine-tuned on annotated datasets to classify or extract adverse

event mentions from noisy social media content. RoBERTa, in particular, has shown strong performance on benchmark datasets in biomedical NLP tasks, including adverse drug event extraction (Khademi et al. 2024). However, PLMs often require extensive domain-specific training data and lack the reasoning flexibility and prompt-based adaptability of LLMs.

Recently, LLMs such as LLaMA 3.1 (Grattafiori et al. 2024), GPT-3 (Brown et al. 2020), and GPT-4 (Achiam et al. 2023) have demonstrated strong few-shot and zero-shot capabilities through in-context learning. In this study, we applied both zero-shot and few-shot learning strategies in combination with CoT prompting, and fine-tuned the LLaMA 3.1-8B-Instruct model on an annotated binary classification dataset to identify Reddit posts mentioning adverse reactions to shingles vaccines. The model achieved robust performance, with precision, recall, and F1-score all exceeding 95% on both the validation and test datasets. This work is conducted as part of the SMM4H-HeaRD 2025 shared task, specifically Task 6, which focuses on identifying personal adverse reactions to shingles vaccines from social media data.

Reddit Post	Label
“ Shingles Vaccine: I had my 1st one in December. Felt like I had the flu for a day and a half afterwards. ”	Positive
“Should I get Shingrix at 22?: I didn’t need prior auth, just a prescription sent in by my doctor... I spend all day every day calling pharmacies for other people, so I’m good at dealing with the same stuff in my own life.”	Negative

Table 1: Examples of adverse events for the shingles vaccine in social media posts.

Method

Our approach utilizes the LLaMA 3.1 model. Each Reddit post was processed by augmenting it with prompts, and the model was subsequently fine-tuned using these prompt-enhanced inputs. The dataset, fine-tuning configurations,

and prompt design are described in detail in the following subsections.

Dataset

This task is framed as a binary classification problem: posts that mention an adverse vaccine event are labeled positive, and all others are labeled negative. The dataset was collected using the Python Reddit API Wrapper (PRAW) by searching combinations of vaccine-related terms associated with shingles (herpes zoster) (Khademi et al. 2024). Posts were manually labeled by two domain experts. A post was labeled positive (“1”) only if it met all three criteria simultaneously: (1) a **personal experience** is described, (2) the post includes an **adverse reaction or symptom**, and (3) the reaction is **specifically related to a shingles vaccine**. Posts that did not meet all criteria were labeled negative (“0”).

Examples illustrating adverse event annotations are presented in Table 1. The positive sample explicitly describes a **personal experience** with **flu-like symptoms** after receiving the shingles vaccine. Conversely, although the negative sample is related to the shingles vaccine, it does not mention any **personal adverse reaction** or **symptoms following vaccination**. Table 3 summarizes the dataset, which comprises Reddit posts annotated for the presence of personal adverse reactions to shingles vaccines. The dataset is split into 2,521 training posts, 786 validation posts, and 8,106 test posts, each containing roughly equal proportions of positive and negative examples.

Methods	P(%)	R(%)	F1(%)
Validation dataset performance			
TweetEval	94.2	97.5	95.8
Zero-Shot (ZS)	96.6	93.7	95.2
Few-Shot (FS)	95.6	95.4	95.5
Chain-of-Thought (CoT)	97.2	95.4	96.3
FS + CoT	97.5	94.3	95.8
Test dataset performance			
Chain-of-Thought (CoT)	94.9	96.6	95.7

Table 2: Performance for validation and test datasets using diverse prompt strategies.

Supervised Fine-Tune Strategy

We fine-tuned the Llama 3.1-8B-Instruct model using the following hyperparameters: a cutoff length of 2048 tokens, a learning rate of 5×10^{-5} , a batch size of 2, and 12 training epochs. The training was conducted on a Linux-based GPU node with three NVIDIA A100 PCIe GPUs (40 GB each) on the Lonestar6 system at the Texas Advanced Computing Center.

Prompt Engineering

This study evaluates the effectiveness of LLMs in detecting personal health mentions related to shingles vaccines, utilizing various prompt engineering strategies—including a Non-CoT prompt (presented in the Supplementary Document) and a CoT prompt designed to guide model reasoning—under both zero-shot and few-shot learning settings.

For few-shot prompting, we included 2 in-context examples illustrating clear personal adverse reactions as described in the supplementary. An example CoT prompt is shown below:

Zero-shot learning CoT prompt

Analyze the following Reddit post and determine if it contains a personal account of an adverse reaction to the shingles vaccine. To decide, follow these steps:

1. Check for first-hand experience cues (e.g., first-person pronouns such as “I”, “my”, or specific references indicating personal experience).
2. Look for an explicit statement that the individual received the shingles vaccine.
3. Identify any mention of side effects or adverse reactions following the vaccination.
4. Evaluate the context: if the language is ambiguous or merely speculative, lean towards a classification of ‘0’.

After considering these points, provide your answer as:

- ‘1’ if the post clearly includes a personal account of adverse reactions.
- ‘0’ if it does not.

Finally, include a brief explanation of your reasoning. Now, classify the following post accordingly.

Result

The dataset was benchmarked using TweetEval (Barbieri et al. 2020), a RoBERTa-based model fine-tuned specifically for the social media domain, and LLMs fine-tuned using four distinct prompt strategies. Table 2 summarizes the performance of these prompting strategies on the validation and test datasets, reporting precision (P), recall (R), and F1-score (F1) specifically for posts labeled positive, which mention adverse vaccine reactions.

On the validation set, the F1-scores for all prompt strategies with the LLM ranged from 95.2% to 96.3%. The CoT approach using zero-shot learning achieved the highest performance, with an F1-score of 96.3% on the validation dataset and 95.7% on the test dataset, demonstrating its effectiveness and robustness in model reasoning. Additionally, the TweetEval model achieved considerable performance, yielding the second-best F1-score of 95.8%. This indicates that traditional domain-specific PLMs remain highly effective for this task while requiring fewer computational resources and providing higher efficiency.

Acknowledgments

The authors acknowledge the department of Research Computing Services at The University of North Texas for providing High Performance Computing resources that have contributed to the research results reported within this paper. URL: <https://research.unt.edu/research-services/research-computing>.

References

- Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Antypas, D.; Ushio, A.; Barbieri, F.; Neves, L.; Rezaee, K.; Espinosa-Anke, L.; Pei, J.; and Camacho-Collados, J. 2023. Supertweeteval: A challenging, unified and heterogeneous benchmark for social media nlp research. *arXiv preprint arXiv:2310.14757*.
- Barbieri, F.; Camacho-Collados, J.; Espinosa Anke, L.; and Neves, L. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In Cohn, T.; He, Y.; and Liu, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1644–1650. Online: Association for Computational Linguistics.
- Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.
- Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Khademi, S.; Palmer, C.; Dimaguila, G. L.; Javed, M.; and Buttery, J. 2024. Exploring Large Language Models for Detecting Online Vaccine Reactions. In *Health. Innovation. Community: It Starts With Us*, 30–35. IOS Press.
- Kwon, S.; and Park, A. 2023. Examining thematic and emotional differences across Twitter, Reddit, and YouTube: The case of COVID-19 vaccine side effects. *Computers in Human Behavior*, 144: 107734.
- Melton, C. A.; Olusanya, O. A.; Ammar, N.; and Shaban-Nejad, A. 2021. Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: A call to action for strengthening vaccine confidence. *Journal of Infection and Public Health*, 14(10): 1505–1512.

Supplementary Document

Datasets	# Positive	# Negative	# Total
Training	1,149	1,372	2,521
Validation	366	420	786

Table 3: Overview of dataset distribution.

Error Analysis

We identified 18 Reddit posts that were consistently misclassified across all four prompting strategies, in which 13 **false negatives** (FN) and 5 **false positives** (FP). Table 4 lists representative misclassification cases, uncovering several recurring sources of misclassification. Implicit phrasing (e.g. beat me up), vaccine ambiguity (mentioning other vaccines

like pneumonia) or structured symptom listings where key information appears in decontextualized numeric form often lead to false negatives as shown in rows 1-3. Rows 4-6 shows false positives frequently triggered by loss of structural metadata due to the nested comments feature of Reddit posts, or by misleading information such as historical or comparative language which does not have an actual event linked to shingles vaccination.

To address these limitations, we need a more robust modeling of user attribution, discourse structure, and temporal relationships. Table 5 summarizes five recurrent error sources and their proposed fixes. Briefly, posts can (i) drift to other vaccines, (ii) bury symptoms in long narratives, (iii) split key cues across bullet lists or numerals, (iv) lose speaker attribution when comments are flattened, or (v) mislead with historical or comparison statements. To address these issues, we suggest exploring streamlined approaches—vaccine-type tagging, span-level symptom extraction, synthetic list augmentation, author-aware segmentation, and temporal/discourse cues—which may enhance robustness in future work.

Non-CoT Few-Shot Learning Prompt Example

Few-Shot Learning w/o Prompt

Decide if a Reddit post is a personal account of adverse vaccine reactions (specifically for the shingles vaccine). A personal adverse reaction mention means someone is describing their own or another individual's side effects after the shot. If so, answer '1'; if not, answer '0'.

Examples:

- Post: 'Just got my second shingles vaccine and wow, what a ride! The first one was a breeze, but this time I had all the bells and whistles of side effects. Fingers crossed that it's all worth it in the end!' → 1.
- Post: 'What a Coincidence! Pushing the Shingles Jab now! Shingles one of the known side effects of the COVID Jab.' → 0.

Now classify the following post accordingly.

#	Type	Post Examples (truncated)	Trigger
1	FN	“Turned 50 and got my first shingles vaccine yesterday... it beat me up but the 2nd one was way easier...”	Phrase dilution
2	FN	“Had my pneumonia vaccine done today... the anticipation was way worse than the actual shot...”	Vaccine ambiguity
3	FN	“Shingrix vaccine symptoms for 7 days... 60F, 5’6”, Asian, Fever 101F, Muscle aches, extreme tiredness... ”	List formatting / numbers
4	FP	“Question: Bleeding after 2nd shingles vaccine and medication changes.: Yeah. I think that might be it. I’m going to wait a couple days.”	Loss of hierarchical or structural metadata of Reddit post
5	FP	“PSA: Vaccinated? Yes, you CAN still get shingles... Yeah, my first outbreak after my vaccine was about a year later... outbreaks... it sucks.”	Misleading info, historical Shingle outbreak not about the vaccine
6	FP	“Got my J&J today... side-effects mild compared to Shingrix shots. ”	Misleading info, comparison sentence contains two vaccine names

Table 4: Misclassified examples from the vaccine adverse reaction classifier.

Trigger	Observation	Potential Solution
Vaccine ambiguity	Some posts primarily mention COVID or other vaccines, causing shingles-related indicators to become absent.	Add vaccine-type tags in preprocessing (e.g., [SHINGLES], [COVID]).
Phrase dilution	Personal reactions are sometimes embedded within longer narratives, making them challenging to detect explicitly.	sentence-level or span-based classification approaches, specifically highlighting text spans containing first-person pronouns adjacent to symptom mentions.
List formatting / numbers	Bullet lists split “101 F”, “muscle aches” across tokens; model trained on sentence prose.	Fine-tune on synthetic bullet-list data; treat “°F/°C” and numerals as fever markers.
Loss of hierarchical or structural metadata	Context formed by joining post and comments with “:”, but key info (personal experience, AE, shingles vaccine) may come from different users, causing attribution ambiguity	Track speaker turns and segment by user to preserve attribution; use structured parsing or speaker-aware models.
Misleading information	Mentions of pre-vaccine illness misinterpreted as post-vaccine AR, or comparisons between vaccines misread as personal reactions.	Use temporal cues and comparison markers to distinguish event timing and intent; apply discourse-aware models.

Table 5: Common misclassified reasons and potential solutions in vaccine-related posts.