School of Computer Science and Engineering

Faculty of Engineering

The University of New South Wales

Vaccine Stance Classification on Social Media

by

Arvind Chandrasekaran

Thesis submitted as a requirement for the degree of

Bachelor of Engineering in Software Engineering

Submitted: 26 November 2025

Supervisor: Dr Aditya Joshi and Dr Jiaojiao Jiang

Student ID: z5503831

Abstract

Understanding public stance towards vaccination is essential for monitoring misinformation and guiding effective public health communication, particularly as social media has become a major platform for the expression and shaping of opinion.

This study aims to compare the performance of large language models (LLMs) — both encoder-only (BERT and RoBERTa variants) and decoder-only (GPT variants) — across multiple optimisation methods, including prompt engineering, few-shot learning, and fine-tuning for inferring stance towards vaccination using posts from the social media platform Reddit.

A custom human-labelled Reddit dataset of vaccine-related posts was developed for both training and evaluation. Key findings show that fine-tuned decoder-only models outperform encoder-only models; specifically, the `gpt-4.1-2025-04-14` variant achieved the highest macro-averaged F1 Score of 95.4%, clearly indicating superior accuracy in vaccine stance classification.

Further application of a RoBERTa variant to a large-scale Reddit dataset covering 2008 to 2025 revealed fluctuations in vaccine stance linked to notable sociopolitical events, offering evidence that stance shifts correspond with external factors.

Acknowledgements

I want to thank my supervisors, Dr Aditya Joshi and Dr Jiaojiao Jiang, for their expertise and erudition, fast feedback and responses, and continuous vital advice and ideas, without whom the thesis would not exist.

Also, I would like to thank Dr Flora Salim, my assessor, for her invaluable feedback and advice.

Abbreviations

LLM Large Language Model

BERT Bidirectional Encoder Representations from Transformers

GPT Generative Pre-trained Transformer

NLP Natural Language Processing

ML Machine Learning

RoBERTa Robustly Optimized BERT Pretraining Approach

DL Deep Learning

BoW Bag of Words

TF-IDF Term Frequency-Inverse Document Frequency

SVM Support Vector Machine

CNN Convolutional Neural Network

LSTM Long Short-Term Memory

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The COVID-19 pandemic highlighted the importance of public health interventions, such as vaccination, in mitigating a mass health crisis. It also underscored that, in the modern digital era, public sentiment towards health interventions like vaccination can fluctuate easily, primarily due to misinformation, disinformation, and concerns that spread online. These fluctuations in vaccine sentiment made public immunisation programs challenging and disrupted continuous efforts to end the pandemic.

Moreover, peer-reviewed research and media reports have shown that in 2020 and 2022, vaccine hesitancy in the United States directly correlated with socio-political factors, with Republican counties having lower vaccination rates than Democratic counties, as shown in Figure 1.1 [1], [2]. These dynamics demonstrate that anti-vaccine campaigns can strategically target specific socio-political groups, sometimes merging with their political identities, making public immunisation even more challenging.

Therefore, understanding vaccine stance online — attitude, opinion, or position of an online media towards vaccination — is essential as uncovering this stance will enable the public health agencies to specially observe the anti-vaccine media online, and hence allowing them to uncover and address any vaccine-related socio-political trends, concerns, hesitation, and misinformation before they translate into real-world consequences. Thus, detecting vaccine stance online is important even in the post-COVID

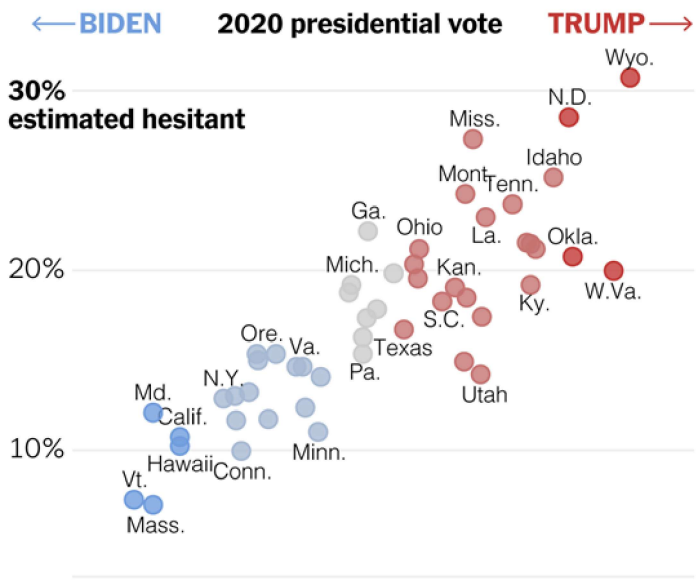

Figure 1.1: Vaccine Hesitancy by U.S. Counties in 2020.

era, as it prepares us for a public immunisation campaign, should it be required in the future.

Additionally, social media provides valuable data for trend analysis. As people consider social media a means to raise awareness and draw the attention of elected representatives, its content is an effective reflection of public opinion on various subjects, including vaccination [3]. Hence, regular monitoring of vaccine stance on social media can help detect anti-vaccine trends or misinformation at an early stage, allowing for timely intervention.

The existing literature on vaccine stance analysis compares different Natural Language Processing (NLP) techniques, ranging from Machine Learning (ML) to Deep Learning (DL) and Transformer-based models. However, the existing literature on this topic does not provide a comparison of different encoder-only and decoder-only Language Models with an evaluation of prompt engineering and fine-tuning approaches on a human-labelled dataset specifically for vaccine stance. The literature also does not provide a pipeline for applying these NLP models to observe vaccine-related trends on social media.

In this research, we address this gap by comparing different Large Language Models (LLMs) using prompt engineering and fine-tuning approaches and evaluating their performance on a custom human-labelled Reddit dataset for vaccine stance inference. Moreover, we applied a transformer-based model to a separate Reddit dataset and observed trends related to vaccine stance, providing a model pipeline for future applications in the public health sector.

Chapter 2 explains the problem of vaccine stance inference and discusses the existing methods, while elucidating the gap in the literature.

Chapter 3 discusses the Large Language Models (LLMs) and the custom dataset used in this research, and also explains the prompt-engineering methods used in this comparison.

Chapter 4 covers the experimentation and the results of this comparative analysis.

Chapter 5 showcases the application pipeline and the observation of trends.

Chapter 6 concludes this research with directions for future work.

Chapter 2

Background

In this chapter, we provide an introduction to the background knowledge for this research. We detail the primary objective of this research — inferring vaccine stance from social media posts. We also discuss the existing literature on this topic, explaining the different methods used for this analysis, while also highlighting the gap in this literature.

2.1 Vaccine Stance

Vaccine stance is defined as a person's overall attitude, opinion, or position towards vaccination. It represents how a person thinks and feels about vaccines and predicts their behaviour towards them.

A person's vaccine stance usually reflects their likelihood of getting vaccinated themselves in the future, as well as ensuring that their dependents, such as children or elderly family members, receive recommended vaccines. Because of this, vaccine stance is widely used in research to understand and anticipate vaccination behaviours in a population.

In research, vaccine stance is widely classified into three sentiment classes: positive,

negative, and neutral.

The **positive sentiment label** represents all posts that support vaccination or show a genuine interest in getting vaccinated. An example from Reddit would be:

> *Just booked my appointment for the booster shot! Feeling relieved and grateful we have access to vaccines.*

This sentiment class also contains posts that highlight the safety of vaccination with the intention of encouraging others. For example:

> *Got my second dose yesterday. Mild soreness.. nothing serious. It feels good to know I'm protecting my family too.*

The positive sentiment label even includes posts that discourage anti-vaccine discourse or posts that condemn the anti-vaccine community for their actions. An example would be:

> *Sad to see that a significant portion of the Western population is distrustful of vaccines.*

The **negative sentiment label** represents all posts that are opposite to the positive label. This includes posts that oppose vaccination or show a lack of interest in getting vaccinated. An example from Reddit would be:

> *I'm not getting another dose. I don't trust how fast these vaccines were pushed out, and I don't want it in my body.*

In this example, the author had already received the first dose of the vaccine at the time of writing the post, but because they refused to take the second dose, the post is

classified as negative. This is because stance reflects the likelihood of getting vaccinated in the future, rather than past vaccination decisions.

The negative class also contains posts that share unpleasant or negative experiences of getting vaccinated with the intention of discouraging others from getting vaccinated. For example:

> *Got my second shot yesterday and I've been completely wiped out... fever, chills, body aches. Remember this before taking your shot.*

Lastly, the negative class also includes posts that expressed hostility towards the vaccine-supporting community. A shortened example from Reddit is:

> *Tbh feels like some of the vaxxed just want everyone else to get it too, like if side effects pop up later we all suffer together.*

The last sentiment class used in this classification is the **neutral label**. This label handles all posts that do not express a vaccine stance. These include posts where the author shares their experience of getting vaccinated without making any remarks that discourage or encourage others. This does not mean that the author does not have a stance. Instead, it means that the stance is not reflected in the post. The neutral label also handles posts that are unrelated to this analysis, but accidentally appear in the dataset. These are posts without any original content; posts that simply share links to other posts and news articles.

Examples from Reddit include:

- *My workplace scheduled our vaccination slots for next week.*
- *Got my second dose today. Had to wait about 30 minutes, and my arm is a little sore now.* This example shared a negative experience, but was considered neutral. Since it was not possible to determine whether the author intended to discourage

others, the post may simply have been sharing a personal experience and was therefore classified as neutral.

2.2 Dataset

In this research, we used social media posts on vaccination from Reddit to perform this vaccine stance classification. Our focus was on textual posts that contained original content, rather than posts that simply shared links to other posts or media articles. For example, a post that only shared a news article, like *"Check out this article about vaccine safety: [link]"*, would be excluded.

We analysed the vaccine stance expressed by a person in each individual post. This means that if the same author made multiple posts, each post was evaluated independently and was assigned its own stance label, based solely on the content expressed in that specific post.

Moreover, the stance label represented the post author's stance for that individual post and not any other person's stance that may have been mentioned. For example, *"My parents just refuse to get vaccinated and say that vaccines cause autism. I'm honestly fed up."* In this post, the author describes their parents' negative sentiment towards vaccination, but they themselves express a positive sentiment towards it, and hence, these kinds of posts were considered positive in our analysis.

2.3 Challenges

We categorise the challenges associated with vaccine stance classification of a social media post into two broad categories.

2.3.1 Public Health Domain Knowledge

This category includes challenges that arise due to the specialised public health domain knowledge required to interpret vaccine-related content. Vaccine-related posts frequently use technical terms related to vaccine types, vaccine makers and even the side effects caused by vaccination.

- Specific vaccines such as *Sputnik* and *Covishield*, and even vaccine-makers' names such as *Pfizer* and *Moderna*, may be used instead of "vaccine", and hence the NLP technique used for this classification must be able to interpret these terms.
- Other terms, such as *jab*, *booster*, and *shot*, can be used interchangeably with "vaccine", and hence, the NLP technique used for this classification must have a great deal of knowledge on these public health terms.
- Vaccine-related discourses often feature different side effects like *arm pain*, *headache* and even terms like *autism*, and hence the selected NLP techniques must be able to interpret these terms along with the associated sentiment. For example, the NLP technique must be able to differentiate *"The vaccine caused light arm pain, not much of a problem"* (positive) from *"The vaccine caused autism, do not take them"* (negative).

2.3.2 Social Media Language

This category encompasses challenges that are rooted in the linguistic features of social media communication. Posts are often informal, abbreviated, and context-dependent, which complicates accurate stance inference.

- Vaccine-related posts can be sarcastic, and this sarcasm can occur to different extents. Higher levels of sarcasm may be undetectable from the post itself and will require additional context about the author. For example, the post *"Another booster shot, yaayyy"* could be a sarcastic post by a vaccine sceptic or could

be a medical professional expressing genuine happiness after having to wait for government legalities for approval. Since this research was focused on the text body of the post, we ignored this higher level of sarcasm. There are other lower levels of sarcasm detectable through the text, and the selected NLP technique should be able to detect these. For example, *"Vaccines are indeed the best way to get autism, yippee"* (negative).

- Social media posts can be expected to have abbreviations, informal vocabulary, typographical errors, and grammatical errors. A short example for this would be — *"Sputnik is goated imo for high effcacy"* (positive). This form of language can affect the performance of NLP techniques, and the technique must be familiar with informal vocabulary words, like *"goated"*, abbreviations like *"imo"* (abbreviation for "in my opinion"), and typographical errors like *"effcacy"* (representing efficacy) to be able to infer the sentiment.

2.4 Vaccine Stance Classification Methods

In this section, we compare the different methods of performing vaccine stance classification on social media by analysing the existing literature on this topic. We will also identify any gaps in this literature.

2.4.1 NLP Techniques for Sentiment Analysis

Natural Language Processing (NLP) techniques to perform sentiment analysis can be classified into four major categories based on their evolution [4]:

1. Lexicon-based
2. Machine Learning (ML)
3. Deep Learning (DL)
4. Transformer-based

Lexicon-based

The Lexicon-based approach is a rule-based method to infer the sentiment of a piece of text. This approach is based on the assumption that the collective sentiment of a text is the aggregation of the sentiment of each individual word within it.

These methods have a pre-defined index or dictionary (*lexicon*) of words, where each word is assigned a sentiment value. The algorithm matches words in the input text against this dictionary. It does not "learn" from data; it simply looks up values.

Most of these lexicon-based methods assign a sentiment score from -1 to +1 to each word in their dictionary, with -1 representing extremely negative sentiment and +1 representing extremely positive sentiment. A mathematical model, such as summation, is used to aggregate the individual sentiment scores of all the words in a text to produce a final sentiment score for the text.

Some of the models take intensifiers like *very* and *extremely* into account by increasing the sentiment magnitude of the words that follow them.

Some common Lexicon-based approaches:

- *SentiWordNet*: A lexical resource assigning sentiment scores to WordNet synsets [5].
- *VADER*: A rule-based tool specifically tuned for social media sentiment [6].
- *TextBlob*: A library used for simple polarity and subjectivity calculations [7].

Machine Learning

The Machine Learning approach is a statistical way of inferring sentiment.

It works by first vectorising words into numerical representations called vectors. Then it finds decision boundaries (lines or hyperplanes) that best separate vectors belonging to one sentiment class from those belonging to the remaining two sentiment classes.

Some common vectorising models:

- *Bag of Words (BoW)*: BoW represents a text as a collection of word frequencies, without considering grammar or word order. It creates a feature space where each dimension corresponds to a unique word in the corpus [4].
- *Term Frequency-Inverse Document Frequency (TF-IDF)*: A more advanced statistical method that measures the importance of a word to a document in comparison to its importance in the entire corpus. It combines term frequency (TF) with inverse document frequency (IDF) to downweight common words (like "the") and highlight rare, discriminative terms [4].

Some common classification models:

- *Support Vector Machine (SVM)*: The model finds the decision boundary by maximising the mutual distance between the boundary and the two closest points from different sentiments [4].
- *Random Forest*: The Random Forest model constructs multiple decision trees and determines the final decision boundary by aggregating the majority vote across these trees, which reduces overfitting and improves generalisation performance [4].

Deep Learning

Deep Learning (DL) models work with vector representations of a text, similar to the ML models, but denser.

Some common classification models:

- *Long Short-Term Memory (LSTM)*: LSTM networks model sentiment by processing a piece of text sequentially and learning how earlier words influence later

ones; they use memory cells and gating mechanisms to retain or forget information, enabling them to capture long-range dependencies in sentences [4].

- *Convolutional Neural Networks (CNNs)*: CNNs classify sentiment by applying filters over sequences of word vectors to detect local patterns such as key phrases. This is followed by pooling operations to highlight the most informative features for the final prediction [4].

Machine Learning and Deep Learning models have shown superior performance to lexicon-based methods in most of the sentiment analysis use-cases [4].

Transformer-based

State-of-the-art in sentiment analysis, transformer-based models process the entire text simultaneously. They utilise self-attention to weigh the significance of each word in relation to every other word in the sentence, capturing deep contextual nuances.

Transformer-based models have shown superior performance over other methods in many recent studies [4].

Transformer-based models can be classified into two classes: *Encoder-only* and *Decoder-only*, based on their architectures [8].

Encoder-only models, such as *BERT* and *RoBERTa*, are built entirely from the encoder stack proposed in the original Transformer architecture. Research shows that these models use bidirectional self-attention, enabling them to capture context from all tokens simultaneously rather than in a left-to-right or right-to-left manner. This deep, bidirectional contextualisation makes encoder-only models particularly effective for natural language understanding (NLU) tasks such as sentiment analysis and classification [8].

In contrast, *decoder-only* models, such as *GPT*-based architectures, rely on the decoder stack and operate using causal (unidirectional) self-attention. This means each token

attends only to previous tokens, enabling the model to generate text in a left-to-right autoregressive manner. Research highlights that this structure is essential for tasks involving text generation, including dialogue, summarisation, reasoning, and open-ended question answering. Decoder-only models excel at generative tasks because they learn strong next-token prediction capabilities during large-scale pre-training [8].

Both encoder-only and decoder-only models demonstrate high performance without the requirement of any further training. This performance is primarily linked to the pre-training of these models on massive corpora using objectives such as masked language modelling. The quality and type of the data vary by language model, and consequently, these models show different performance for the same task [8].

Both encoder-only and decoder-only models can be fine-tuned, which involves taking a pre-trained transformer model and updating all or part of its parameters on a smaller, task-specific dataset. This allows the model to adapt its representations to domain-specific patterns and typically yields strong performance for tasks like sentiment analysis [8].

Moreover, decoder-only models, due to their generative nature, can be *prompt-engineered*. Prompt engineering does not modify model weights. Instead, it guides a pre-trained model using natural-language instructions or examples. In *zero-shot prompting*, only an instruction is provided; in *few-shot prompting*, a few labelled examples are included to demonstrate the task [8].

Conclusion

In conclusion, the review highlights that transformer-based models tend to outperform traditional ML and DL models in most of the recent studies of sentiment analysis. ML and DL trail Transformers in performance, followed by lexicon-based methods.

2.4.2 Empirical Evaluation of ML, DL and Transformer-based Methods

Research from 2023 compared different vaccine stance classification methods, ranging from Machine Learning models such as SVM, Naïve Bayes, and Decision Tree to Deep Learning Models such as LSTM and CNN. Some of these comparisons have even included BERT.

Jain et al. (2023) compared Machine Learning techniques with LSTM and BERT, using COVID-19 vaccine opinions from X (previously known as Twitter) [9].

The study sourced an initial dataset of 10,000 COVID-19 vaccine opinion tweets from Kaggle, collected using an undisclosed methodology. These tweets were then labelled using a heuristic-based lexicon approach called TextBlob.

To improve data consistency, the researchers then removed all the neutral tweets, resulting in a refined dataset of 3,700 tweets with only positive and negative sentiments.

This final set was balanced, containing roughly 2,000 positive and 1,700 negative tweets, and was split into an 80:20 ratio for training and testing.

The authors employed five ML algorithms: *Support Vector Machine (SVM)*, *Naïve Bayes*, *Logistic Regression*, *Decision Tree*, and *Random Forest*. Vectorisation was done using *Bag of Words (BoW)* and *TF-IDF*.

The study also utilised *LSTM* as one Deep Learning method for this comparative analysis, and *BERT* as the only transformer-based approach.

The results showed that the best performing ML model was SVM with TF-IDF vector representations, with an accuracy of 88.79%. LSTM and BERT reported an accuracy of 88.26% and 90.42% respectively.

In conclusion, BERT came out as the best-performing technique, which was followed by SVM and LSTM. This difference between BERT and the second-best-performing method, SVM, was less than 2%, which is not very significant.

Alkhushayni et al. (2023) conducted a comparative study of machine learning (ML) and deep learning (DL) approaches to analyse public sentiment regarding the COVID-19 vaccine on X [10].

The study used the Twitter API to collect English-language tweets using keywords such as #Covid19vaccine #CoronaVaccine. The ground-truth labelling for this dataset was done using two lexicon-based sentiment analysers: *TextBlob* and *Vader*. The data was also preprocessed to remove noise (hashtags, URLs, emojis). The dataset was finally split into a 75:25 ratio for training and testing.

The authors employed two classical ML algorithms: *Support Vector Machine (SVM)* and *Decision Tree* (using both Gini and Entropy criteria), utilising Unigram, Bigram, and 4-gram feature extraction methods.

They also used three deep learning architectures built using *Keras*: a *Densely Connected Neural Network*, a *Convolutional Neural Network (CNN)*, and a *Long Short-Term Memory (LSTM)* network, for comparison.

The results demonstrated that:

- *LSTM* was the best-performing model overall, achieving a high accuracy of 95.26% followed by *CNN* with an accuracy of 91.03%.
- Amongst ML models, the *Decision Tree* (using Entropy and Unigram/Bigrams) outperformed *SVM*, reaching 87% accuracy compared to SVM's 86%.

In conclusion, this study shows that classical ML methods were inferior in comparison to DL methods such as *LSTM* and *CNN*, for vaccine stance inference, having accuracies of 95.26% and 91.03% respectively, a significant jump from that of Jain et al. (2023).

Ahmed et al. (2023) conducted a sentiment analysis regarding five specific COVID-19 vaccines: AstraZeneca, Moderna, Pfizer/BioNTech, Sinopharm, and Sputnik. This study aimed to establish a framework to compare the performance of Machine Learning (ML), Deep Learning (DL) models and *BERT* [11].

The study used a dataset of approximately 20,967 tweets. To ensure high-quality ground truth, the data was manually labelled as 'in favour' or 'against' by three graduate students, with a label assigned only if at least two of them agreed.

The Machine Learning model used in this research included: *Random Forest*, *Extra Tree Classifier*, and *Naïve Bayes*. Vectorisation of the input data was done using *Word2Vec*, *TF-IDF* and *BoW*. These Machine Learning models were compared to the Deep Learning models of *CNN* and *LSTM*. Lastly, *BERT* was also used in this comparative analysis.

The results demonstrated something unique. We observed that *BERT* reported an accuracy of 85% which was more than other DL methods used in this analysis, which had an accuracy close to 80%. However, the ML algorithm of *Extra Tree Classifier* outperformed *BERT* with an accuracy of 92% when used with *BoW*.

In conclusion, we observe that the research from 2023 shows that Transformer-based models are a superior sentiment analysis technique over Deep Learning and Machine Learning on average. However, in one study from 2023, we observe that the *Extra Tree Classifier* with *BoW* vectors outperformed *BERT* by a margin of 7%. In all the cases, Transformer-based models reported consistent high performance with an accuracy of more than 85%. This shows that Transformer-based models are consistently better than ML and DL models. Therefore, we choose to focus on Transformer-based models for this research and will evaluate their performance through our experimentation.

2.4.3 Empirical Evaluation within Transformer-based Models

Recent research from 2024 and 2025 has compared the Transformer-based models for similar public health-related sentiment analysis tasks on a human-labelled dataset. We analyse this research and observe any gaps within the methodology in each case.

Lossio-Ventura et al. (2024) performed a comparative analysis of sentiment analysis tools on a health survey data related to COVID-19. The study compared eight widely

used state-of-the-art tools and investigated the performance of Large Language Models (LLMs) via few-shot and zero-shot learning approaches in this specific domain [12].

The study sourced two independent datasets comprising survey responses regarding COVID-19 experiences collected by the National Institutes of Health (NIH) and Stanford University during the pandemic lockdown.

The NIH dataset (26,411 sentences) focused on mental health impacts, while the Stanford dataset (21,266 sentences) captured positive effects ("silver linings"), difficulties, and reasons for isolation.

A subset of 1,260 sentences was manually labelled by human annotators on a 3-point scale (negative, neutral, positive). This subset was split into a training set of 260 sentences, which was used for few-shot learning, and a test set of 1,000 sentences. The data was labelled by three independent labellers to ensure quality, with final labels determined by majority agreement.

The study compared multiple NLP techniques:

- **Lexicon-based approaches:**
  - LIWC2015: Calculates word frequencies across psycholinguistic categories.
  - SentiStrength: Uses linguistic rules and a lexicon of terms with polarity intensity.
  - TextBlob: A library utilising a rule-based model and Naïve Bayes classifier.
  - VADER: A rule-based model optimised for social media text.
- **Deep Learning Models:**
  - Stanza: A Convolutional Neural Network (CNN) model trained on word vectors and syntactic features.
- **Transformer (Encoder-only):**
  - TweetEval: A RoBERTa model fine-tuned on tweet sentiment analysis.

  - Pysentimiento: Based on BERTweet (a RoBERTa model) and fine-tuned on English tweets.
  - NLPTown: A BERT-based multilingual model fine-tuned on product reviews.

- **Transformer (Decoder-only):**
  - OPT (Few-Shot): Open Pre-Trained Transformers (1.3B and 2.7B parameters)
  - ChatGPT (Zero-Shot): GPT-3.5 relies on its pre-training knowledge.

The results showed that fine-tuned encoder-only models like TweetEval and Pysentimiento outperformed lexicon-based methods, and the decoder-only LLMs demonstrated superior performance. The zero-shot ChatGPT approach achieved the highest performance, outperforming the few-shot OPT models by approximately 6% in accuracy and 4–7% in F1 score across both datasets.

In conclusion, this study demonstrates that Transformer-based models are superior in performance to other approaches, with decoder-only models outperforming encoder-only. However, this study fails to utilise decoder-only LLMs to the best of their ability. This study lacks a comparison between zero-shot and few-shot prompt-engineering methods on the same model and fine-tuning.

Parsa and Dubey (2024) evaluated the performance of Large Language Models (LLMs) for analysing public sentiment toward vaccines on social media. The study aimed to provide a system for public health monitoring by introducing a novel dataset and comparing the efficacy of open-weight versus closed-source models [13].

The study collected data from the Reddit community `r/Vaccine`, covering the period from June 9, 2019, to June 9, 2024. A total of 3,110 posts and comments were filtered to remove metadata and irrelevant entries, resulting in a final dataset of 970 vaccine-related messages. This dataset was manually labelled as positive, neutral, or negative to serve as the training and test data for the evaluation.

The study focused exclusively on Transformer-based models. The models used in this analysis were:

- **Encoder-only models:** FacebookAI's RoBERTa Small: A Small Language Model (SLM) that was fine-tuned on the dataset. The authors used a 75:25 training-test set split.
- **Decoder-only models:** GPT-3.5 Turbo, GPT-4o, Llama3-70B, and were used without any fine-tuning.

Ensemble Methods included:

- **Majority:** Determines sentiment based on the majority opinion of the three LLMs (GPT-3.5, GPT-4o, Llama3); if there is no agreement, it returns "neutral".
- **Override:** Defaults to the prediction of the best-performing individual model (Llama3-70B) unless the other two models reach a consensus, in which case the consensus overrides.

The results indicated that the Llama3-70B model was the most accurate individual model (85.1%), outperforming the closed-source GPT-4o (79.8%) and GPT-3.5 Turbo (71.8%). The ensemble methods provided marginal improvements, with the "Override" method achieving the highest overall accuracy of 86.9%. The fine-tuned RoBERTa model achieved an accuracy of 72.0%.

Although the study is a good measure for vaccine stance classification, it fails to demonstrate a comparison between fine-tuned, few-shot prompt-engineered and zero-shot prompt-engineered methods for decoder-only LLMs. The study does not utilise decoder-only LLMs to their best potential. Moreover, this study provides a model for inferring the vaccine stance; however, it fails to demonstrate how this vaccine stance classification can be used in real life with an application pipeline and correlation with real-world events.

Gowda et al. (2025) conducted a comparative study to evaluate the effectiveness of different ML and Transformer-based models on multiple datasets from different domains. We will be focusing on the sentiment analysis within the healthcare domain [14].

The study utilised a specific healthcare dataset comprising 30,000 patient reviews and discussions. The data was labelled with sentiment towards treatments and providers, divided into three classes of positive, negative and neutral. With a significant class imbalance ratio of 1:1:5, negative sentiment being too dominant, the researchers employed data augmentation techniques, specifically using back-translation to generate synthetic positive samples.

For the healthcare domain, the following Transformer models were fine-tuned:

- **Encoder-only:** BERT-base-uncased, RoBERTa base
- **Decoder-only:** GPT (version undisclosed)

The results demonstrated that RoBERTa achieved the highest performance in the healthcare domain, with an accuracy of 94% and an F1-score of 93%. This was followed by GPT, which achieved 93% accuracy. BERT, which achieved 92% accuracy and 91% F1-score, trailed both of these models.

In conclusion, the study compares BERT-base-uncased, RoBERTa base and an undisclosed version of GPT for healthcare. However, it fails to explore prompt-engineering methods like zero-shot and few-shot learning and synthetically augments the dataset, making the analysis less reliable for our use case.

To summarise, we observed that the existing literature on the public health domain fails to satisfy our requirements of vaccine stance classification on social media. We observed that the literature that compares Transformer-based models on a human-labelled vaccine stance dataset fails to employ the different methods of prompting and fine-tuning a decoder-only LLM. Moreover, it fails to provide an application pipeline

that will enable the public health industry to use vaccine stance by correlating it with real-world events. Additionally, other public health-related literature failed to perform a complete gold-standard analysis with a comparison of fine-tuning and prompting methods on a human-labelled dataset without any synthetic augmentations. We address this gap with our research.

2.4.4 GPT LLMs

Previously, we observed that GPT-3.5 with a zero-shot prompt outperformed OPT despite having few-shot examples [12]. We note that an LLM's pre-trained knowledge plays a significant role in determining its performance in an inference task, like sentiment analysis. This depends on the quality of the data on which these LLMs have been trained.

Ahmad et al. (2025) introduced VaxGuard, a novel dataset and framework designed to evaluate the detection of Large Language Model (LLM)-generated vaccine misinformation. The study aimed to identify the LLMs most effective at detecting such content across diverse spreader roles and narratives [15].

The researchers employed a zero-shot, prompt-based detection methodology to evaluate: GPT-3.5, GPT-4o, LLaMA3, PHI3, and Mistral. The models were provided with a prompt asking them to classify input text into binary categories: "Otherwise" (0) or "Misinformation" (1) without any fine-tuning.

The results demonstrated the dominance of GPT-3.5 with an F1 score of 96%.

In conclusion, GPT models are likely trained on more data related to vaccination than the other LLMs involved in this comparison, making them ideal for any task that requires vaccine-related domain knowledge. Hence, in our research, we decided to invest our resources in GPT models.

Chapter 3

Method

As discussed in the previous section, Transformer-based Large Language Models (LLMs) are the best-performing stance classification technique for the given use of vaccine stance and social media dataset. However, the existing research fails to compare different types of LLMs — including encoder-only and decoder-only — with different prompt engineering and fine-tuning approaches on a human-labelled dataset from Reddit. In this chapter, we discuss the dataset, LLMs, and prompt-engineering and fine-tuning approaches used in our research. These methods will be compared in the next chapter, thereby addressing this gap in the literature.

3.1 Training and Test Dataset

To ensure that the training and evaluation of the chosen LLMs were effective, we created a custom human-labelled dataset dedicated to this vaccine stance inference. This dataset consisted of Reddit posts collected from a publicly available, unlabelled Reddit dataset on Kaggle [16].

This Kaggle dataset contains over 60,000 English-language Reddit posts from more than 600 subreddits, having content up to October 25, 2021.

The dataset features diverse ideologies and opinions, which is essential for effective training and evaluation. The posts included content from mainstream vaccine discussion communities, such as *r/AskDocs*, as well as other lesser-known communities, for public health discourses, like *r/conspiracy*. This ensured that training data was diverse and not just saturated with discussions from subreddits dedicated to public health discourses.

The primary reason for using this dataset was the availability of posts from the banned subreddit of *r/NoNewNormal*. This subreddit was started by the vaccine-sceptic community to discuss the supposed harmful effects of vaccination and was the most popular amongst the anti-vaccine communities on Reddit during the pandemic. This community became so influential in spreading vaccine-related disinformation and misinformation that it was banned by the U.S. government and subsequently removed from Reddit with all of its content [17]. The inclusion of *r/NoNewNormal* provides additional coverage of anti-vaccine discourse, which is unavailable in many new datasets.

To ensure diversity in terms of discourse and relevance to vaccination, the dataset was filtered using the following list of keywords: *vaccine, vaccination, immunization, booster, vaxx, vax, vaxxed, vaxxer, unvaxxed, vaccine safety, vaccine efficacy, vaccine hesitancy, pro-vax, anti-vax, COVID vaccine, Moderna, Pfizer, AstraZeneca, Johnson & Johnson, mRNA vaccine, vaccine passport, vaccine mandate, vaccine rollout, vaccine equity, MMR, HPV vaccine, flu shot, polio vaccine, DTaP, and Tdap.* These keywords were selected after evaluating vaccine-related discourses online on Reddit and other platforms.

Once filtered, up to 1502 posts from 512 unique subreddits were manually evaluated for relevance and were labelled as one of the three stances: positive, negative, or neutral. We were able to manually label 500 posts from the positive and negative sentiment classes and 502 posts from the neutral sentiment class.

Subsequently, the labelled dataset was split into training (90%) and test sets (10%), with the training set comprising a total of 1350 posts with 450 posts from each of the three stance labels, and the test set comprising a total of 152 posts with 50 posts from

the positive and negative classes and 52 posts from the neutral class.

3.2 Large Language Models

The Transformer-based language models used in this research can be classified based on their architecture as encoder-only or decoder-only.

In the previous chapter, within the literature review, we observed that OpenAI's GPT models were the best-performing LLMs for vaccine-related tasks. Hence, we decided to use multiple GPT-based models to represent decoder-only models in this comparative analysis.

The decoder-only models included were:

1. gpt-4.1-2025-04-14 [18]: with a context window of 1,047,576 tokens and knowledge cut-off up to 1st June 2024, gpt-4.1-2025-04-14 was the largest (undisclosed parameter size) and well-pretrained GPT model released by OpenAI at the time of the experiment [18].
2. gpt-4.1-mini-2025-04-14 [19]: similar to gpt-4.1-2025-04-14 in terms of pre-training but smaller.
3. gpt-4.1-nano-2025-04-14 [20]: smallest version of gpt-4.1-2025-04-14.
4. gpt-4o-2024-08-06 [21]: with a context window of 128,000 tokens and knowledge cut-off up to 1st October 2023.
5. gpt-3.5-turbo-0125 [22]: legacy model of OpenAI, it had a context window of 16,385 tokens and a knowledge cut-off up to 1st September 2021.

All these models were accessible through OpenAI's API [23].

In the previous chapter, we also observed that BERT-based and RoBERTa-based models were amongst the best-performing encoder-only LLMs with accuracies over 85%.

Hence, the encoder-only models included in our research were BERT and RoBERTa-based.

The encoder-only models included were:

1. Google's BERT-base-uncased [24]: the smallest version of BERT with a parameter size of 110M. This BERT model is an English language model that was pre-trained on BooksCorpus [25], a dataset consisting of 11,038 unpublished books, and English Wikipedia [26] (excluding lists, tables, and headers).

2. Google's BERT-large-uncased [24]: a larger variant of the same BERT with a parameter size of 340M. This BERT model is also an English language model trained on the same dataset as BERT-base-uncased.

3. FacebookAI's RoBERTa large [27]: RoBERTa large has a parameter size of 355M and was pretrained on a combined 160GB English corpus consisting of Book-Corpus (11,038 books), English Wikipedia (excluding lists, tables, and headers), CC-News (63 million articles from Sept 2016–Feb 2019), OpenWebText (an open-source recreation of GPT-2's WebText), and Stories (CommonCrawl data filtered for story-like Winograd-style text).

4. Cardiff NLP's twitter-roberta-base-sentiment-latest [28]: RoBERTa large variant fine-tuned on a Twitter dataset for sentiment analysis.

All these model weights were available on the Hugging Face Hub and accessible through the Hugging Face Transformers library [29].

3.3 Baseline Zero-Shot Prompt Engineering

The first method used in this research was a simple zero-shot baseline prompt that explained the three sentiment classes of vaccine stance.

The prompt is structured into three points, with each point explaining one of the classes. This was followed by the instruction to classify a given post into one of the three explained classes. No additional information or instructions were provided. Moreover, this prompt was zero-shot and hence no examples were provided within the prompt.

The purpose of this zero-shot baseline prompt is to compare its performance with other few-shot and optimised methods explored in this research and observe any improvements.

This baseline zero-shot prompt was provided as a system instruction using the OpenAI API. While the posts themselves were passed as user instructions, one after another. The temperature parameter of all the LLMs was set to zero to ensure deterministic and predictable behaviour.

The exact prompt for this vaccine stance inference:

"The positive stance includes posts that support vaccination, show genuine interest in taking it, encourage others, discourage negative sentiment towards vaccination, and counter the anti-vaccine community.
The negative stance includes posts that oppose vaccination, lack interest in getting vaccinated, discourage others from taking the vaccine, and counter the vaccine community.
The neutral stance consists of posts that are not clearly positive or negative. Posts can share experiences (can be painful), but the intention is not to discourage vaccination.

You are an expert in vaccine stance inference. Using the above information, infer the stance of the following Reddit post."

3.4 Role-Based Incremental Coaching (RBIC)

The Role-Based Incremental Coaching (RBIC) approach used in this research was a modified version of the research by Ding et al. (2024) [30].

RBIC is a zero-shot prompt engineering framework that effectively leverages a decoder-only LLM's chain-of-thought capability.

RBIC works in two phases. The first phase is the knowledge-generation phase, where the language model is prompted to generate all the pre-requisite information required for the task. This pre-generation of all the necessary information reduces the hallucination that may happen when the model is directly prompted to do a task. With the information being pre-generated, the model will take it into the context along with the subsequent prompt, allowing for a better output than if the subsequent prompt were provided on its own.

The second phase is the incremental coaching phase, which involves dividing a task into smaller and simpler steps and tackling them sequentially. This division into smaller steps makes each of the incremental coaching prompts simpler, reducing mistakes.

In the case of vaccine stance, within the knowledge-generation phase, the language model was provided with an explanation of the three stance labels — similar to the zero-shot baseline prompt engineering method — as a system instruction. The model was also prompted to generate some linguistic cues that might be present in the posts; all of this can be useful in the next phase.

In the incremental coaching phase, the first step involved providing the language model with the post from which the stance was supposed to be inferred. In the second step, the language model was prompted to simplify the language used in the provided social media post without changing its meaning. In the last step, the language model was prompted to return a stance label for this simplified post, using all the provided and generated knowledge.

RBIC is a zero-shot approach, and thus, examples were not provided in the prompt. The temperature parameter of all the LLMs was set to zero to ensure deterministic and predictable behaviour. The exact prompt used in this research is presented in Table 3.1.

Role	**Prompt**	**Output**	**Phase**
system	*"You are an expert in inferring a post's stance towards vaccination. Vaccine Stance is of 3 types - Positive: Posts that support vaccination, show genuine interest in taking it, encourage others, discourage negative sentiment towards vaccination, and counter the anti-vaccine community. Negative: Posts that oppose vaccination, lack interest in getting vaccinated, discourage others from taking the vaccine, and counter the vaccine community. Neutral: Posts that are not clearly Positive or Negative. It could be sharing experience (can be painful) but the intention is not to discourage vaccination. "*	None	Knowledge-Generation
user	*"List some common linguistic cues present in textual Reddit posts that contain vaccine sentiment that can help infer the sentiment of the post."*	Linguistic cues	Knowledge-Generation
assistant	Linguistic cues	None	Incremental Coaching
user	*"{Post} Simplify the language of the post without changing its meaning. Capture the key points and express them in a clear and concise manner."*	Simplified post	Incremental Coaching
assistant	Linguistic cues and the simplified post	None	Incremental Coaching
user	*"Infer the post's stance towards vaccination, with all the information. (Return only the sentiment (Positive/Negative/Neutral))"*	Stance label	Incremental Coaching

Table 3.1: RBIC Prompt

3.5 DSPy Few-Shot Prompt Engineering

This is the first method that employs the large training set, which was manually labelled for this research. Declarative Self-Improving Python (DSPy) is a framework that enables us to optimise the instructions — in the form of a prompt — given to an LLM [31],[32].

The DSPy optimises a simple instruction to a few-shot prompt using two tools, namely, the ChainOfThought module and the BootstrapFewShotWithRandomSearch optmizer (spelt as in the DSPy documentation).

3.5.1 ChainOfThought Module

The ChainOfThought (CoT) module is a framework to structure a prompt in an efficient JSON-like object. The prompt describes the input and output variables, along with explaining the process of obtaining the expected output from a given input. Additionally, the model adds a reasoning value to the output variables, forcing an LLM to return a reasoning value along with the necessary output for the task. This reasoning value makes an LLM think and reduce hallucination.

In the case of vaccine stance inference, the exact way the prompt — provided as a system instruction — was structured is as follows:

"Your input fields are:

1. 'post' (str): Social media post from which the poster's stance towards vaccination needs to be inferred.Your output fields are:

1. 'reasoning' (str):

2. 'stance' (Literal['positive', 'negative', 'neutral']): One-worded stance label.All interactions will be structured in the following way, with the appropriate values filled in.

[[# post #]]

{post}[[# reasoning #]]

{reasoning}[[# stance #]]

{stance} # note: the value you produce must exactly match (no extra characters) one of: positive; negative; neutral[[# completed #]]

In adhering to this structure, your objective is:

The stance towards vaccination is of 3 types:

1. positive: Posts that support vaccination, show genuine interest in taking it, encourage others, discourage negative sentiment towards vaccination, and counter the anti-vaccine community.
2. negative: Posts that oppose vaccination, lack interest in getting vaccinated, discourage others from taking the vaccine, and counter the vaccine community.
3. neutral: Posts that are not clearly positive or negative. It could be sharing experience (can be painful) but the intention is not to discourage vaccination. "

3.5.2 BootstrapFewShotWithRandomSearch Optimizer

DSPy presents an optimizer called BootstrapFewShotWithRandomSearch that selects a small list of effective few-shot examples — to be added to the above ChainOfThought (CoT) module prompt — from a large labelled training set such that the few-shot examples maximise the stance classification performance on a separate validation set. This is achieved by selecting t random candidate sets of n examples each from the labelled training dataset and then evaluating each of these candidate sets using the CoT module's standard prompt on a separate validation set. For every candidate set, the optimizer also runs the base CoT module on the training examples to identify which examples produce the correct output. From these, it selects k random examples that not only have the right answer but also include the reasoning steps generated by the base CoT model. These reasoning traces are then stored along with the final outputs, allowing the few-shot examples to pass both the answer and the reasoning process into

the optimised prompt. The candidate set that achieves the highest performance on the validation data is then chosen as the final few-shot configuration. Although the DSPy framework operates with a linear metric, such as accuracy, that can be applied to every individual entry in the validation set and then summed up towards the end to maximise, in this research, we maximised for the macro-averaged F1-score instead, which is a nonlinear metric. We achieved this by implementing the changes in our own version of the DSPy code.The training set, containing 1350 posts (450 posts from each of the three labels), was further subdivided into a smaller training and validation set. The validation set accounted for 20% of this training set and contained 270 posts (90 posts from each of the three stance labels). Lastly, the hyperparameter values of t, n, and k were 16, 4, and 16, respectively, and the temperature parameter was set to zero to ensure deterministic behaviour.

3.6 Fine-tuning

The last method used in this research was fine-tuning. The entire training dataset, comprising 1350 posts (450 posts from each of the three stance labels), was used to fine-tune.

3.6.1 Decoder-only

All the decoder-only LLMs belonged to OpenAI's GPT suite and hence were fine-tuned using OpenAI API's supervised fine-tuning approach [33]. Supervised fine-tuning trains the model on labelled input–output pairs, where the model learns to produce the target output by minimising a supervised loss. The decoder-only LLMs were fine-tuned using the following hyperparameters: batch size = 2, learning rate multiplier = 2, epochs = 3, and temperature = 0. The learning rate is a fixed value, which has not been disclosed by OpenAI. The hyperparameters were automatically selected by the API and were determined based on the size of the training dataset [34].

3.6.2 Encoder-only

All the encoder-only LLMs were fine-tuned using Low-Rank Adaptation (LoRA). Cardiff NLP's Twitter-roberta-base-sentiment was already fine-tuned on a X (previously known as Twitter) dataset. It was further fine-tuned for this vaccine stance inference use case. LoRA is a parameter-efficient fine-tuning method that updates only a small set of low-rank matrices inserted into the model, instead of modifying all model weights. This significantly reduces the computational cost and memory requirements of training while still allowing the model to learn task-specific behaviour [35]. The following LoRA configurations and hyperparameters were used:

LoRA Configurations:

- Low-rank dimension = 8
- LoRA alpha = 16
- Target modules = ["query", "value"]
- LoRA dropout = 0.1 (used for regularisation)

Hyperparameters:

- Learning rate = $3 * 10^{-5}$
- Per-device training batch size = 16
- Weight decay = 0.01
- Temperature = 0

Chapter 4

Experimentations and Results

In this chapter, we explain the experimentation setup, including the test dataset and the evaluation metric. Additionally, we discuss the results and compare the LLMs and methods explained in Chapter 3.

4.1 Experimentation Setup

4.1.1 Test Dataset

The test dataset consisted of a total of 152 human-labelled posts, with 50 positive, 50 negative, and 52 neutral posts.

4.1.2 Evaluation Metrics

Since we are performing a three-class stance classification, a model's performance is unlikely to be identical for all three classes. A linear metric like accuracy can be biased in its reflection of performance. Instead, we use the macro-averaged F1 score.

F1 Score per class

The F1 score for each class is computed as the harmonic mean of precision and recall:

- Precision measures how many of the posts predicted as that class were actually correct.
- Recall measures how many posts belonging to that class were correctly identified by the model.

$$Precision = TruePositives/(TruePositives + FalsePositives)$$

$$Recall = TruePositives/(TruePositives + FalseNegatives)$$

$$F1Score = 2 * Precision * Recall/(Precision + Recall)$$

Macro-Averaged F1 Score
Unweighted average of the individual F1 scores

4.1.3 Experimentation

All the Large Language Models (LLMs), including all the encoder-only and decoder-only models, were fine-tuned on the training dataset, and their performances were evaluated and compared on the test dataset.
Based on the macro-averaged F1 score, the best-performing decoder-only model was selected for the zero-shot and few-shot prompt engineering methods discussed previously.
All the zero-shot and few-shot prompt-engineering methods were used with this selected decoder-only model and applied to the test dataset 5 times. The fine-tuned model was run 5 times on the same test dataset. The macro-averaged F1 scores were averaged for these 5 runs, and the variance was also calculated.

Model	**Architecture**	**Macro-Avg-F1**
gpt-4.1-2025-04-14	Decoder-only	95.4%
gpt-4.1-mini-2025-04-14	Decoder-only	94.09%
gpt-3.5-turbo-0125	Decoder-only	93.98%
gpt-4o-2024-08-06	Decoder-only	92.07%
gpt-4.1-nano-2025-04-14	Decoder-only	88.67%
cardiffnlp/twitter-roberta-base-sentiment	Encoder-only	85.9%
RoBERTa-large	Encoder-only	85.24%
BERT-base-uncased	Encoder-only	80.68%
BERT-large-uncased	Encoder-only	79.02%

Table 4.1: Comparative Analysis of Fine-tuned Models

4.2 Results and Observation

4.2.1 Comparison of all the LLMs

Table 4.1 presents the macro-averaged F1 score of all nine fine-tuned encoder-only and decoder-only models used in this research. The table lists these models in decreasing order of their performance.

Overall Performance Trends

We observe in Table 4.1 that decoder-only GPT models significantly outperformed encoder-only models on this vaccine-stance inference task. The overall top-performing model, gpt-4.1-2025-04-14, achieved a macro-averaged F1 score of 95.4%, indicating extremely strong performance across all stance categories. In contrast, the top-performing encoder-only model, twitter-roberta-base-sentiment, achieved 85.9%, which is considerably lower than the GPT family models.

We can infer that decoder-only GPT models — with their rich and diverse training data — were better equipped at handling the sarcastic and informal communication style of social media, paired with the required domain knowledge related to vaccination. Encoder-only BERT-based and RoBERTa-based models lacked this intense pre-training and hence showed much lower performance.

Performance Across Decoder-Only GPT Models

Additionally, we observed a consistent performance hierarchy within the decoder-only models that correlated with the model's parameter count (size of the model) and pre-training quality:

- gpt-4.1-2025-04-14 (95.4%) was the biggest model with the best pre-training in our analysis. Thus, it demonstrated the highest capability in understanding informal language of social media discourse.
- gpt-4.1-mini-2025-04-14 (94.09%) and gpt-3.5-turbo-0125 (93.98%) are smaller variants that also performed well, but with a drop of $\tilde{1}$.4% to 1.5% from the biggest model.
- gpt-4o-2024-08-06 (92.07%) belongs to the Omni series. Being slightly older, this model has inferior pre-training quality. Thus, it was slightly behind the above generation of GPT-4.1-based models.
- gpt-4.1-nano-2025-04-14 (88.67%) is a smaller model designed for efficiency rather than accuracy, and hence had he lowest performance amongst the GPT models.

Performance Across Encoder-Only Models

We observed that CardiffNLP's twitter-roberta-base-sentiment was the best-performing encoder-only language model, which was fine-tuned twice: first for sentiment analysis on a dataset from X (previously known as Twitter) and second on a vaccine stance inference dataset from Reddit.
Both the RoBERTa-based models ranked higher than the BERT-based models. Within the BERT-based models, BERT-base-uncased (110M parameters) ranked higher than BERT-large-uncased (340M parameters) despite being significantly smaller, with an improvement of 1.66%. This was unlike what we observed in the case of decoder-only models, where pre-training quality and parameter count were more influential.

Method	**Average and Variance of 5 Macro-Avg-F1**
Fine-tuning	95.4 and 2.4e-5
Zero-Shot(DSPy ChainofThought alone)	91.33 and 0
Few-Shot(DSPy BootstrapFew-ShotWithRandomSearch)	91.33 and 0
Zero-Shot(RBIC)	90.99 and 1.23e-32
Zero-Shot(Baseline)	90.99 and 4.0e-05

Table 4.2: Comparative Analysis of Prompt Engineering Methods on gpt-4.1-2025-04-14

4.2.2 Comparison of the Prompt Engineering Methods

Table 4.2 presents the prompt engineering methods, which were discussed in Chapter 3, in decreasing order of performance. The best-performing decoder-only model was selected for this comparison with the aim of observing the improvement of these prompt engineering methods over the baseline, while additionally comparing them with fine-tuning.

Performance

In terms of performance, we observed that fine-tuning the language gives the best performance with a macro-averaged F1 score of 95.4% with a very low variance of 2.4e-5 across five runs. This indicates that a task-specific fine-tuned model is not only highly accurate but also consistently reliable.

The fine-tuned model was followed by the zero-shot DSPy ChainofThought module applied independently, and the few-shot DSPy BootstrapFewShotWithRandomSearch approach on top of the ChainofThought module. Both approaches had the same macro-averaged F1 score of 91.33% with literally a variance of 0 for the five runs. This shows that the few-shot examples had negligible influence on the performance of the ChainOfThought module for the selected test set. Fine-tuning improved the performance of the ChainOfThought module by 4.07% which is significant.

The role-based incremental coaching (RBIC) had a negligible improvement over the baseline prompt, with both having the same macro-averaged F1 score of 90.99% and

variances of 1.23e-32 and 4.0e-05, respectively.

Variance

Fine-tuning showed a small but non-zero variance of 2.4e-5. This shows that the model had minor fluctuations across runs despite having zero as the value for temperature. In contrast, the zero-shot DSPy ChainofThought and few-shot DSPy BootstrapFewShotWithRandomSearch methods showed zero variance, indicating highly deterministic behaviour in their inference process. This consistency can be attributed to several factors. First, the use of a JSON-like structure for prompting allowed for a highly organised specification of the inputs and the expected outputs, which reduced ambiguity in model interpretation. Second, the prompts also included a description of how to derive the output from the input. Finally, the inclusion of reasoning values as part of the output further constrained the model, effectively preventing random hallucinated responses. This ensured that gpt-4.1-2025-04-14 produced consistent and repeatable outputs for the same set of inputs, despite the fact that absolute determinism cannot be guaranteed even when the temperature parameter is set to zero.

Chapter 5

Application

In this chapter, we will discuss the application of the best-performing encoder-only model — twitter-roberta-base-sentiment — to a separate dataset from Reddit, performing temporal analysis of trends while observing any correlation with real-world events that might influence vaccine discourses online.

5.1 Dataset

Apart from the training and test dataset discussed previously, the application was done on a separate Reddit dataset that was provided to us by Epiwatch, UNSW Kirby Institute [36].

This dataset comprised a total of 11,202 posts ranging from April 25, 2008, to October 24, 2025 and was collected from Reddit using an undisclosed list of keywords from 16 different subreddits: *r/Health, r/medicine, r/publichealth, r/nursing, r/Vaccines, r/doctors, r/AskDocs, r/China_Flu, r/Coronavirus, r/CoronavirusUK, r/COVID19, r/CoronavirusUS, r/CoronavirusAustralia, r/CoronavirusDownunder, r/CoronavirusCanada, and r/LockdownSkepticism.*

All of these subreddits are mainstream platforms for vaccine discussions and are non-

partisan communities without any inclinations towards a specific stance. Some of the subreddits, like *r/LockdownSkepticism* and *r/China_Flu*, may have names that sound partisan; however, these communities were found to be non-partisan after evaluating their subreddit descriptions [37],[38]. Hence, posts from all these subreddits can be expected to fall under any of the three stance labels.

The posts were made from 27 different countries, with the USA (36.65%), India (21.72%), Canada (15.21%), the UK (6.92%), Australia (6.79%) and Italy (5.31%) accounting for up to 92.6% of all the posts.

5.2 Application Pipeline

In this section, we discuss the application pipeline used in our research to perform a temporal analysis of trends in this dataset, detailing the processes used in each step.

5.2.1 Cleansing

The first step in our analysis was removing entries that were irrelevant. We created a filter to remove all the posts that did not have a text body. However, in this case, none of the posts were removed.

5.2.2 Clustering

The second — and most important — step of our analysis was to identify the different vaccine-related discourse topics present in the dataset. These discourse topics represent the various aspects of vaccination that people talk about in their posts. A topic could focus on a specific vaccine manufacturer, or it could involve broader correlations discussed by users, such as "vaccines and autism".

Our aim in this step was to identify all these topics and group together all posts from the dataset that were associated with each topic, essentially clustering them into topic-

based groups called clusters. Once we have these clusters, we can observe the trends of positive, negative and neutral sentiments in each of these clusters over time, thus performing a temporal analysis of trends.

To achieve this, we employ the following two methods. Firstly, a keyword-based clustering approach. In this method, we compiled a list of popular and well-known vaccine-related topics, along with a set of potential keywords that are expected to appear in posts discussing each of these topics. We then used these keywords to identify and group together posts.

In our research, we used the following list of topics and keywords, which were found after performing a manual evaluation of posts from the training dataset discussed in Chapter 3:

1. Vaccine Types

We defined topics based on major vaccine types and included common variations of their names:

- Flu: “flu”, “influenza”
- Measles: “measles”, “rubeola”, “mmr”
- COVID-19: “covid”, “covid-19”, “sars-cov-2”, “coronavirus”
- Chickenpox: “chickenpox”
- Rabies: “rabies”

2. Vaccine Brands / Manufacturers

To capture discussions about specific vaccine manufacturers, we created brand-based topics with their associated spellings and variants:

- Pfizer: “pfizer”, “biontech”

- Moderna: "moderna"
- AstraZeneca: "astrazeneca", "astra zeneca"
- Janssen (J&J): "janssen", "johnson & johnson"
- Novavax: "novavax"
- Sinovac: "sinovac"
- Sinopharm: "sinopharm"
- Sputnik V: "sputnik v"
- Covaxin (Bharat Biotech): "bharat biotech", "covaxin"

3. Symptoms and Anxiety

We included two thematic categories to differentiate between posts discussing physical symptoms and posts expressing emotional states:

- Symptom-related keywords: "fever", "chills", "headache", "pain", "ache", "fatigue", "rash", "nausea", "cough"
- Anxiety-related keywords: "anxiety", "anxious", "nervous", "fear", "worried", "panic", "stress"

4. Vaccination Status

These keywords helped identify posts describing whether users were vaccinated or unvaccinated:

- Vaccinated phrases: "got vaccinated", "was vaccinated", "received vaccine", "fully vaccinated"
- Unvaccinated phrases: "not vaccinated", "unvaccinated", "anti-vax", "refuse vaccine"

5. Dose Phase

To differentiate between posts referencing different stages of the vaccination process, we added:

- First dose: "first dose", "1st dose", "initial dose"
- Second dose: "second dose", "2nd dose"
- Booster dose: "booster", "3rd dose", "third dose", "additional dose"

Since the keyword-based approach did not cover all the possible topics and required manual evaluation, we used an unsupervised approach to create more clusters — BERTopic Modelling [39]. BERTopic modelling works in the following four steps:

1. Embedding the Posts: Each post is first converted into a numerical representation, called an embedding, which captures the semantic meaning of the text. These embeddings are created in such a way that similar posts have similar representations, even if they use different words. In this research, we utilised the multilingual MiniLM model [40] to generate these embeddings, which ensured that non-English posts were not misplaced in the subsequent steps.
2. Dimensionality Reduction: Since embeddings are high-dimensional, BERTopic reduced their dimensions using UMAP. This makes it easier to group similar posts together without losing important semantic information.
3. Clustering: The reduced embeddings are clustered using a density-based algorithm (HDBSCAN). Posts that are close together in the embedding space are assigned to the same cluster, forming topic groups.
4. Topic Representation: For each cluster, BERTopic identifies representative words or phrases that best describe the posts in that cluster. This creates an interpretable topic label, which helps us understand what each cluster is about.

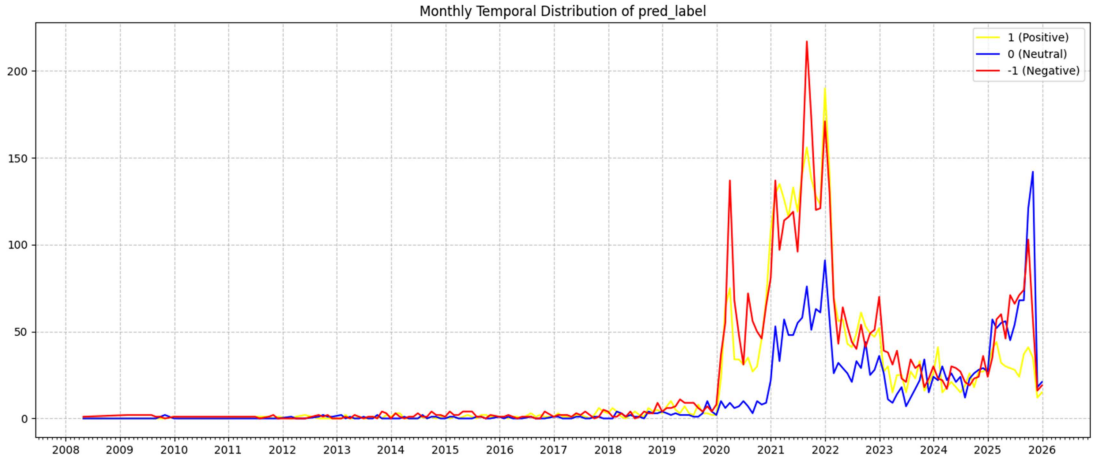

Figure 5.1: Monthly Temporal Distribution of Stance Labels

BERTopic modelling created up to 149 clusters. Table 5.1 explains some of the relevant clusters selected after manual evaluation.

5.2.3 Vaccine Stance Inference

In the last step, the best-performing encoder-only model — Cardiff NLP's twitter-roberta-base-sentiment — was applied to all the posts sequentially and the stance labels were plotted in a graph for trend analysis. The application of an English-language encoder-only LLM was decided based on computational budget constraints.

5.3 Observation

Figure 5.1 shows the monthly temporal distribution of vaccine stance labels predicted by the LLM. The positive, negative, and neutral sentiments are represented by the colours yellow, red, and blue, respectively.

In the temporal distribution of sentiment, we observe that the amount of vaccine discussions on Reddit remained significantly low and stagnated until the start of the COVID-19 pandemic in early 2020. From early 2020, the number of vaccine discussions increased rapidly, with positive and negative sentiments competing with one

another, while the neutral sentiment trailed behind them. The vaccine discussions reached an all-time peak during late 2021 when the public immunisation campaign against COVID-19 began. With the start of 2023, we see a decline in vaccine discussions as we move into the post-COVID-19 era, with all the sentiments having nearly equal discussions.

However, we see a sudden jump again in vaccine discussions starting from 2025. Moreover, the discussion in 2025 is different from what we observed during the COVID-19 pandemic. In 2025, the negative and neutral sentiments have an edge over the positive sentiment, which did not experience the same growth as the other two sentiment classes since 2024.

We investigate this sentiment in the clusters we obtained and observe that within the BERTopic cluster, 44_autism_rfk_jr_acetaminophen (*'autism', 'rfk', 'jr', 'acetaminophen', 'vaccines', 'claims', 'tylenol', 'cdc', 'link', 'cause')* is one of the biggest negative clusters, being the most popular in June and September of 2025.

This increase in negative discussion correlates with many real-world events, including the start of Donald Trump's second presidency in 2025, and the appointment of Robert F. Kennedy Jr., a well-known vaccine sceptic, as the United States Secretary of Health and Human Services. Additionally, September 2025 also witnessed the American President Donald Trump linking Tylenol with Autism and subsequently banning it [41]. This action by Donald Trump triggered vaccine-related discussions on Reddit, particularly with negative sentiment.

This correlation was also stated by Alemi & Lee (2023), where vaccine hesitancy was linked to socio-political inclinations. Republican counties in the US were the least vaccinated.

Cluster Name	**Keywords**	**Number of Posts**	**Detail**
44_autism_rfk_jr_acetaminophen	*'autism', 'rfk', 'jr', 'acetaminophen', 'vaccines', 'claims', 'tylenol', 'cdc', 'link', 'cause'*	46	This cluster contained posts that discussed about Robert F. Kennedy Jr., a well-known vaccine skeptic and his claims on vaccination causing autism.
41_pregnancy_pregnant_during_maternal	*'pregnancy', 'pregnant', 'during', 'maternal', 'women', 'infants', 'milk', 'lactating', '19', 'imz'*	48	This cluster contained posts that discussed about getting vaccinated during pregnancy — questions and concerns.
39_hpv_cervical_cancer_vaccine	*'hpv', 'cervical', 'cancer', 'vaccine', 'sexually', 'gardasil', 'smear', 'my', 'pap', 'it'*	50	One of the vaccine types missed in the keyword-based approach. This cluster contained posts discussing cervical cancer vaccine.

Table 5.1: Selected BERTopic Modelling Clusters

Chapter 6

Conclusion

6.1 Conclusion

In this thesis, we performed a comprehensive comparative analysis of Large Language Models (LLMs) for vaccine stance inference using social media data, addressing two major gaps in the existing literature.

1. **Comparative analysis:**

   The existing literature on vaccine stance analysis has compared traditional machine learning and transformer-based models, but it has not evaluated a broad spectrum of LLMs — including both encoder-only and decoder-only architectures. This thesis fills that gap by systematically comparing nine different LLMs using prompt engineering (zero-shot, few-shot, RBIC) and fine-tuning approaches on a custom human-labelled Reddit dataset. Our results demonstrate that fine-tuned decoder-only GPT models significantly outperform encoder-only BERT and RoBERTa models, with gpt-4.1-2025-04-14 achieving a macro-averaged F1 score of 95.4%.

2. **Application pipeline:**

Existing research does not provide a clear application framework for deploying stance-classification models at scale or for analysing long-term trends in vaccine discourse. This thesis addresses this gap by presenting an end-to-end application pipeline consisting of data cleansing, topic clustering using both keyword-based and BERTopic modelling, and stance inference using the best-performing model. Applying this pipeline to a large Reddit dataset (2008–2025), we demonstrated how vaccine stance fluctuates in response to major real-world events — including the COVID-19 pandemic and political developments in 2025. This pipeline validates the practical usefulness of stance classification and also provides a framework for public health agencies and researchers to monitor vaccine stance over time.

6.2 Future Work

This thesis was mainly focused on the NLP side of vaccine stance classification. In the future, the work can be improved in the following ways.

- First, we did not explore author-level analysis. If we map each post back to the author and study their past posts, we can detect long-term behaviour, changes in stance, and even sarcasm more accurately.
- Second, better temporal modelling methods can be added to observe how people's stance evolves over months or years.

Bibliography

[1] F. Alemi and K. H. Lee, "Impact of Political Leaning on COVID-19 Vaccine Hesitancy: A Network-Based Multiple Mediation Analysis," *Cureus*, vol. 15, no. 8, Art. no. e43232, Aug. 2023, doi: 10.7759/cureus.43232.

[2] D. Ivory, L. Leatherby, and R. Gebeloff, "Least vaccinated U.S. counties have something in common: Trump voters," The New York Times. Apr. 17, 2021. [Online]. Available: https://www.nytimes.com/interactive/2021/04/17/us/vaccine-hesitancy-politics.html. [Accessed: Feb. 26, 2025].

[3] Pew Research Center, "Americans think social media can help build movements, but can also be a distraction," Pew Research Center. Sept. 9, 2020. [Online]. Available: https://www.pewresearch.org/short-reads/2020/09/09/americans-think-social-media-can-help-build-movements-but-can-also-be-a-distraction/. [Accessed: Apr. 10, 2025].

[4] M. K. Chandan and S. Mandal, "A comprehensive survey on sentiment analysis: Framework, techniques, and applications," *Comput. Sci. Rev.*, vol. 58, pp. 1–32, Sept. 2025, doi: 10.1016/j.cosrev.2025.100777.

[5] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC '06)*, Genoa, Italy, May 2006, pp. 1–8. Available: https://aclanthology.org/L06-1225/.

[6] C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. Int. AAAI Conf. Web Social Media*, May 2014, pp. 216–225, doi: 10.1609/icwsm.v8i1.14550.

[7] TextBlob Developers, "TextBlob: Simplified Text Processing," *TextBlob 0.19.0 Documentation*, v0.19.0, Jan. 13, 2025. Available: https://textblob.readthedocs.io/en/dev/. Accessed: Nov. 26, 2025

[8] K. S. Kalyan, A. Rajasekharan, and S. Sangeetha, "AMMUS: A Survey of Transformer-based Pretrained Models in Natural Language Processing," *arXiv preprint arXiv:2108.05542*, 2021.

[9] T. Jain, V. K. Verma, A. K. Sharma, B. Saini, N. Purohit, Bhavika, H. Mahdin, M. Ahmad, R. Darman, S.-C. Haw, S. M. Shaharudin, and M. S. Arshad, "Sentiment Analysis on COVID-19 Vaccine Tweets using Machine Learning and Deep Learning Algorithms," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 5, pp. 32–41, 2023, doi: 10.14569/IJACSA.2023.0140504.

[10] S. Alkhushayni, Z. Alomari, and D. Al-Zaleq, "A Sentiment Analysis Study of Twitter Users' Reactions to the COVID-19 Vaccine," in *2023 14th Int. Conf. Inf. Commun. Syst. (ICICS)*, Irbid, Jordan, 2023, pp. 1–6, doi: 10.1109/ICICS60529.2023.10330455.

[11] S. Ahmed, D. Khan, S. Sadiq, M. Umer, F. Shahzad, K. Mahmood, H. Mohsen, and I. Ashraf, "Temporal Analysis and Opinion Dynamics of COVID-19 Vaccination Tweets Using Diverse Feature Engineering Techniques," *PeerJ Comput. Sci.*, vol. 9, Art. no. e1190, 2023, doi: 10.7717/peerj-cs.1190.

[12] J. A. Lossio-Ventura, R. Weger, A. Y. Lee, E. P. Guinee, J. Chung, L. Atlas, E. Linos, and F. Pereira, "A Comparison of ChatGPT and Fine-Tuned Open Pre-Trained Transformers (OPT) Against Widely Used Sentiment Analysis Tools: Sentiment Analysis of COVID-19 Survey Data," *JMIR Ment. Health*, vol. 11, Art. no. e50150, Jan. 2024, doi: 10.2196/50150.

[13] A. Parsa and A. Dubey, "Evaluation of LLMs, BERT, and Ensemble Techniques for Analyzing Online Vaccine Sentiment," in *2024 Conf. on AI, Sci., Eng., and Technol. (AIxSET)*, Laguna Hills, CA, USA, Sep. 30–Oct. 2, 2024, pp. 162–165, doi: 10.1109/AIxSET62544.2024.00029.

[14] K. P. Gowda, R. Porwal, C. Ramesh, and S. Shekhar, "Transformers in Sentiment Analysis: A Paradigm Shift in Deep Learning Research," *J. Inf. Syst. Eng. Manage.*, vol. 10, no. 5s, pp. 262–280, 2025, doi: 10.52783/jisem.v10i5s.612.

[15] S. T. Ahmad, H. Lu, S. Liu, A. Lau, A. Beheshti, M. Dras, and U. Naseem, "VaxGuard: A Multi-Generator, Multi-Type, and Multi-Role Dataset for Detecting LLM-Generated Vaccine Misinformation," *arXiv preprint arXiv:2503.09103*, 2025, doi: 10.48550/arXiv.2503.09103.

[16] Lexyr Inc., "The Reddit COVID Dataset," Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/pavellexyr/the_reddit_covid_dataset. [Accessed: Jun. 02, 2025].

[17] R. Iyengar, "Reddit takes action against groups spreading Covid misinformation," CNN, Sep. 1, 2021. [Online]. Available: https://edition.cnn.com/2021/09/01/tech/reddit-covid-misinformation-ban. [Accessed: Nov. 02, 2025].

[18] OpenAI, "GPT-4.1," OpenAI Platform, Apr. 14, 2025. [Online]. Available: https://platform.openai.com/docs/models/gpt-4.1. [Accessed: Nov. 02, 2025].

[19] OpenAI, "GPT-4.1 mini," OpenAI Platform, Apr. 14, 2025. [Online]. Available: https://platform.openai.com/docs/models/gpt-4.1-mini. [Accessed: Nov. 02, 2025].

[20] OpenAI, "GPT-4.1 nano," OpenAI Platform, Apr. 14, 2025. [Online]. Available: https://platform.openai.com/docs/models/gpt-4.1-nano. [Accessed: Nov. 02, 2025].

[21] OpenAI, "GPT-4o," OpenAI Platform, May 13, 2024. [Online]. Available: https://platform.openai.com/docs/models/gpt-4o. [Accessed: Nov. 02, 2025].

[22] OpenAI, "GPT-3.5 Turbo," OpenAI Platform, 2023. [Online]. Available: https://platform.openai.com/docs/models/gpt-3.5-turbo. [Accessed: Nov. 02, 2025].

[23] OpenAI, "API," OpenAI, 2025. [Online]. Available: https://openai.com/api/. [Accessed: Nov. 02, 2025].

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2018, doi: 10.48550/arXiv.1810.04805.

[25] Y. Zhu et al., "Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books," Project Page. [Online]. Available: https://yknzhu.wixsite.com/mbweb. [Accessed: Nov. 20, 2025]

[26] "English Wikipedia," *Wikipedia, The Free Encyclopedia*. [Online]. Available: https://en.wikipedia.org/wiki/English_Wikipedia. [Accessed: Nov. 26, 2025].

[27] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692*, 2019, doi: 10.48550/arXiv.1907.11692.

[28] D. Loureiro, F. Barbieri, L. Neves, L. Espinosa Anke, and J. Camacho-Collados, "TimeLMs: Diachronic Language Models from Twitter," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics: System Demonstrations*, Dublin, Ireland, May 2022, pp. 251–260, doi: 10.18653/v1/2022.acl-demo.25.

[29] T. Wolf et al., "HuggingFace's Transformers: State-of-the-art Natural Language Processing," *arXiv preprint arXiv:1910.03771*, 2019, doi: 10.48550/arXiv.1910.03771.

[30] X. Ding, B. Carik, U. S. Gunturi, V. Reyna, and E. H. R. Rho, "Leveraging Prompt-Based Large Language Models: Predicting Pandemic Health Decisions and Outcomes Through Social Media Language," in *Proc. 2024 CHI Conf. Hum. Factors Comput. Syst. (CHI '24)*, Honolulu, HI, USA, May 2024, Art. no. 443, pp. 1–20, doi: 10.1145/3613904.3642117.

[31] O. Khattab et al., "DSPy: Compiling Declarative Language Model Calls into State-of-the-Art Pipelines," in *Proc. 12th Int. Conf. Learn. Represent. (ICLR)*, Vienna, Austria, May 2024. [Online]. Available: https://openreview.net/forum?id=sY5N0zY5Od.

[32] O. Khattab et al., "DSPy Documentation," dspy.ai. [Online]. Available: https://dspy.ai/. [Accessed: Nov. 01, 2025].

[33] OpenAI, "Supervised Fine-tuning," OpenAI Platform, 2025. [Online]. Available: https://platform.openai.com/docs/guides/supervised-fine-tuning. [Accessed: Jun. 10, 2025].

[34] OpenAI, "Fine-tuning best practices," OpenAI Platform, 2025. [Online]. Available: https://platform.openai.com/docs/guides/fine-tuning-best-practices. [Accessed: Jun. 10, 2025].

[35] E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9.

[36] Kirby Institute, "EPIWATCH: Prevent the next pandemic with Epidemic Intelligence," Kirby Institute. [Online]. Available: https://www.kirby.unsw.edu.au/research/projects/epiwatch. [Accessed: Nov. 25, 2025].

[37] "r/LockdownSkepticism," Reddit. [Online]. Available: https://www.reddit.com/r/LockdownSkepticism/. [Accessed: Nov. 25, 2025].

[38] "r/China_Flu," Reddit. [Online]. Available: https://www.reddit.com/r/China_Flu/. [Accessed: Nov. 25, 2025].

[39] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv preprint arXiv:2203.05794*, 2022, doi: 10.48550/arXiv.2203.05794.

[40] M. Grootendorst, "BERTopic," *BERTopic Documentation*. [Online]. Available: https://maartengr.github.io/BERTopic/index.html. [Accessed: Jun. 10, 2025].

[41] BBC News, "Trump makes unproven link between autism and Tylenol," BBC News, Sep. 23, 2025. [Online]. Available: https://www.bbc.com/news/articles/cx20d4lr67lo. [Accessed: Nov. 24, 2025].