Amsterdam University Press

COMPUTATIONAL COMMUNICATION RESEARCH 8.1 (2026) 1–31
HTTPS://DOI.ORG/10.5117/CCR2026.1.4.SUN

When More Shots Don't Help: LLM Sensitivity and Variability in Social Media Annotation and Stance Detection of Health Information

Luhang Sun
School of Journalism and Mass Communication, University of Wisconsin-Madison, United States

Varsha Pendyala
Department of Electrical and Computer Engineering, University of Wisconsin-Madison, United States

Yun-Shiuan Chuang
Department of Psychology & Department of Computer Sciences, University of Wisconsin-Madison, United States

Shanglin Yang
Department of Computer Sciences, University of Wisconsin-Madison, United States

Jonathan Feldman
School of Interactive Computing, Georgia Institute of Technology, United States

Andrew Zhao
Institute of People and Technology, Georgia Institute of Technology, United States

Munmun De Choudhury
School of Interactive Computing, Georgia Institute of Technology, United States

Sijia Yang
School of Journalism and Mass Communication, University of Wisconsin-Madison, United States

Dhavan V. Shah
School of Journalism and Mass Communication, University of Wisconsin-Madison, United States

Abstract

This paper leverages large language models (LLMs) to experimentally determine strategies for scaling up social media annotation and stance detection of health information, with HPV vaccine-related tweets as a case study. We examine both conventional fine-tuning and emergent in-context learning methods, systematically varying strategies of prompt engineering and in-context learning across widely used LLMs and their variants (e.g., GPT-4, Mistral, Llama 3, and Flan-UL2). Specifically, we varied prompt template design, shot sampling methods, and shot quantity to detect stance on HPV vaccination. Our findings reveal that (a) in-context learning outperformed fine-tuning in stance detection for HPV vaccine social media content; (b) increasing shot quantity does not necessarily enhance performance across models; (c) stratified sampling often outperforms random sampling, with the performance gap more pronounced in smaller model variants, and (d) LLMs and their variants present differing sensitivity to in-context learning conditions. This study highlights the potential and provides an applicable approach for applying LLMs to research on social media annotation and stance detection of health information.

Keywords: large language models, in-context learning, prompt engineering, fine-tuning, stance detection, social media annotation, machine learning classification, HPV vaccination

Introduction

Computational social science research often involves applying computer-assisted techniques to process large-scale data resources to understand human behaviors on a societal scale (D. Lazer et al., 2009; D. M. Lazer et al., 2020). Researchers have increasingly recommended integrating AI-assisted techniques and LLMs into the workflow of computational social science research, which can reduce significant costs in terms of time, human labor, financial resources, and technical expertise (Törnberg, 2025; Ziems et al., 2024). Indeed, the rise of LLMs heralds a "paradigm shift" in computational social science, especially for applications to repetitive, expertise-laden, and time-consuming tasks, such as latent feature annotation (also referred to as "labeling" or "coding"), which are essential for training supervised machine learning models. These tools have the potential to revolutionize the study of digital conversation and media technologies, especially in research domains such as health communication and message framing analysis, which require high levels of accuracy and interpretive skill within specific contexts to identify linguistic patterns and generate reliable classifications. Given the nascent nature of the AI turn in computational social science research and the variety of options to implement LLMs in a research project, evi-

dence is needed to identify strategies for AI deployment. We tackle this gap by systematically documenting the performance of LLMs while varying several key dimensions for implementation in the context of large-scale stance detection for social media annotation of health information, with HPV vaccine-related tweets as a case study.

Specifically, we focus on prompt engineering, in-context learning, and fine-tuning approaches, which can potentially enhance the performance of general-purpose LLMs for specific message annotation tasks such as stance detection. Within in-context learning, we systematically varied the following dimensions: prompt design, number of shots, and selection strategies, and different LLM models (i.e., GPT-4, Mistral, Llama 3, Flan-UL2). Our study evaluates the performance of different in-context learning design dimensions and fine-tuning approaches across LLMs, discusses the implications of specific practices, and suggests avenues for integrating LLMs in health communication and computational communication science research. The key practices we explore include the details of the prompt message design, randomized or stratified shot sampling methods, zero-shot, few-shot, and many-shot selection quantity, and fine-tuning strategies of adjusting model parameters. Our goal is to provide practical insights with empirical performance evidence that can help structure future research on health information, especially studies of consequential digital conversations on online platforms where stance detection is crucial in an era of heightened misinformation.

The specific context of the study is human papillomavirus (HPV) vaccine discussions on Twitter, selected due to the prevalence of mixed attitudes and vaccine skepticism on social media platforms (Massey et al., 2020). HPV vaccination, given its relatively shorter history, continues to face low uptake (Vraga et al., 2023). Prior research has also documented the politicization of policy support for the HPV vaccine (Saulsberry et al., 2019). The dynamics emphasize the importance of examining how the HPV vaccine is discussed in online spaces, where attitudes are often mixed and skepticism is expressed alongside support. In this context, stance classification is useful to capture whether messages express agreement, opposition, or neutrality toward the HPV vaccine, thereby enabling researchers and stakeholders to understand the linguistic patterns and rhetorical strategies of public discourse on health information disseminated over media technologies. Recognizing challenges of prior supervised machine learning annotation tasks, we aim to demonstrate the potential of LLMs and their variants in detecting HPV vaccine stance effectively, using both in-context learning experimental design and

fine-tuned models.

Generally, our findings suggest that in-context learning configurations outperform their fine-tuning counterparts, with models and their variants presenting different levels of sensitivity to the manipulated in-context learning conditions in the experiment. Larger model variants perform better under the detailed-prompt condition, often outperforming the basic-prompt condition. Increasing shot quantity does not necessarily enhance performance across models. In addition, the stratified sampling condition often outperforms random sampling, with the performance gap becoming more pronounced in smaller model variants. Across tests, we find GPT-4 Turbo (OpenAI's top commercially available model at the time of this analysis) outperforms other frontier large language models on overall performance metrics, with the best results emerging when using a stratified sampling method, as opposed to the random sampling method, peaking with a "few shot" level of six examples, combined with a detailed contextual prompt. These findings may inform future research on health information across different topics or datasets.

Taken together, our findings highlight both the potential and the challenges of integrating LLMs into computational social science research on health information, emphasizing the importance of refining in-context learning conditions, tailoring model-specific adaptations, and incorporating human-in-the-loop interventions. We contribute to the emerging literature (Demszky et al., 2023) on how LLMs can be harnessed effectively to complement human skills in a cooperative fashion in computational social science.

The context of mixed attitudes toward HPV vaccination in social media discourse

Vaccine hesitancy has been a global challenge to public health (Bussink-Voorend et al., 2022), particularly for the vaccines with relatively shorter histories, such as the low-uptake HPV and COVID-19 vaccination (Vraga et al., 2023). Researchers in public health and health communication have endeavored to study the determinants and interventions of the HPV vaccine and other vaccine hesitancy, as well as the linguistic features of social media discussions around vaccination (Chen et al., 2020; Ortiz et al., 2019; Puri et al., 2020). Social media has been viewed as a crucial channel for the dissemination of both positive and negative attitudes toward HPV and other vaccines, as it has become a primary source of health information for many members of the public, which may potentially affect their perceptions and

intentions regarding vaccination (Dunn et al., 2017; Nan & Madden, 2012; Vraga et al., 2023).

Various empirical studies show that negative attitudes toward vaccination (e.g., concerns about vaccine effectiveness and safety), alongside political ideologies and conspiracy theories, are prevalent around discourses of vaccine hesitancy and skepticism on social media (Dhaliwal & Mannion, 2020; Di Domenico et al., 2022; Massey et al., 2020). Conversely, research finds that greater perceived certainty about the scientific evidence for the HPV vaccine is tied to more support for HPV vaccine policies (Saulsberry et al., 2019). On social media, both pro- and anti-vaccine sentiments are often expressed through personal narratives and anecdotes shared by individual users (Massey et al., 2020). To understand how individual users make sense of their beliefs and intentions regarding vaccination, regardless of their attitudes, it is important to examine the actual content that people encounter or express on social media. Recent studies have examined vaccination discourse on social media content at scale by integrating various computer-assisted techniques. For instance, researchers applied a hierarchical machine learning-based sentiment analysis system to annotate mixed attitudes toward HPV vaccination on Twitter, and found positive, negative, and neutral stances each accounted for roughly one-third of the corpus, with most negative tweets concerning vaccine safety (Du et al., 2017). X. Jiang et al. (2021) combined supervised and unsupervised machine learning annotation techniques to understand COVID-19 vaccine discourse and its ideological dimensions. They operationalized vaccine stance by training human coders to annotate latent message features, such as vaccine favorability and distrust, which were then used to train and fine-tune Bidirectional Encoder Representations from Transformers (BERT) models. Building on the approach of combining human annotation with machine learning for stance detection, the present study applies stance detection to HPV vaccination discourse, using it as a critical tool for identifying in favor, neutral, or opposing views in health-related messages.

Social media discourse around HPV vaccination extends beyond the mixed stances to the broader narratives and focal topics raised by online users. In particular, certain topics frequently surface in the HPV vaccine discussion, including reproductive health and fertility concerns, fears related to disease and death, cervical cancer prevention benefits, vaccination advocacy, and personal experiences and anecdotes (Dunn et al., 2017; Surian et al., 2016; H. Zhang et al., 2020). Examining these topic-specific discourses may shed light on health information acquisition behaviors (S. Jiang et al.,

2023) among online users, and also provide contextual cues for annotation and stance detection tasks. Prior studies suggest that accounting for specific topic context can improve the reliability and validity of stance detection, since thematic features often structure how support, opposition, and neutrality toward health information are expressed on social media (Hanley & Durumeric, 2023). In addition, researchers have found that these prevalent topics discussing HPV vaccination may carry both supportive and oppositional stances. For instance, cancer is often framed positively as evidence of the HPV vaccine's protective role against cervical and other HPV-related cancers (H. Zhang et al., 2020), but it also appears in anti-vaccine discourse alleging that the vaccine itself causes cancer (Kornides et al., 2023; Weinzierl & Harabagiu, 2022). Similarly, reproductive health and fertility are discussed on social media both in terms of the benefits of protecting women's reproductive health (Ortiz et al., 2019) and in claims that the vaccine leads to infertility or other reproductive disorders (Smith & Gorski, 2024; Weinzierl & Harabagiu, 2022). Death is also used in a bidirectional way, with some discourse highlighting HPV-related mortality for vaccination promotion, while others alleging deaths caused by the vaccine (Semino et al., 2023; Weinzierl & Harabagiu, 2022). Taken together, incorporating topic classifications into stance detection of HPV vaccine discourse may help capture the nuanced ways in which mixed attitudes toward HPV vaccination are expressed across different thematic frames, and further test the utility of LLMs in stance detection.

Integrating LLMs into text classification tasks

To better understand the vast and diverse social media landscape surrounding vaccine-related discourses, researchers have turned to computational methods, including dictionary-based approaches (Himelboim et al., 2020; King et al., 2023; Wang et al., 2019), unsupervised machine learning clustering approaches (Hwang et al., 2022; X. Jiang et al., 2021), and supervised machine learning approaches (Chuang et al., 2023; Piedrahita-Valdés et al., 2021; Sun et al., 2023). These approaches enable researchers to classify textual content for stance detection, sentiment analysis, and topic features, which serve as foundational elements for further analysis and interpretation. Admittedly, supervised machine learning models offer the advantage of being tailored to specific data characteristics and message features of interest, by training classifiers on their own for specific contexts and purposes. However, this process may require considerable time, human labor, financial resources, and technical expertise, as a conventional supervised

machine learning approach in computational social science often involves many stages, from data collection, data pre-processing, high-volume human annotations, model training, and evaluation.

The recent rise of LLMs in computational social science has the potential to address some of these challenges by enabling more efficient, accurate, and cost-benefit data analysis. Researchers have begun integrating LLMs to streamline feature construction and classification within large-scale datasets across various research contexts (Heseltine & Clemm von Hohenberg, 2024; Tan et al., 2024). For instance, Ziems et al. (2024) evaluate multiple LLMs for social science applications, recommending the integration of LLMs in annotations and generation tasks due to their high efficiency and cost-effectiveness. Likewise, Törnberg (2025) argues that LLMs outperform supervised classifiers and even expert coders in the context of political social media messages.

While the incorporation of LLMs into computational social science research has proven beneficial, due to its low cost, high efficiency, and performance accuracy, it also raises new questions. As the number of LLMs and their customization options increase, it is time to scrutinize and refine prompt engineering (i.e., designing prompt content for task guidance), in-context learning (i.e., embedding examples within prompts to provide context for the task), and fine-tuning techniques (i.e., adjusting model parameters for enhanced performance) to improve performance across different tasks and LLMs (Yao et al., 2023). Specifically, prompt quality is likely to impact model performance, particularly in tasks like classifying social media content that may require nuanced interpretive skills. In addition to a prompt template *per se*, in-context learning, a component of prompt engineering that embeds relevant examples within prompts, also plays a crucial role in enhancing model performance by providing LLMs with contextual information. Moreover, fine-tuning methods, a conventional, yet complementary approach to prompt engineering, offer distinct advantages for integrating LLMs into text classification. Unlike prompt engineering, fine-tuning does not require additional effort to design prompt templates or provide in-context examples, as it customizes the model's parameters directly to the task. We believe that subjecting these different features of the LLM stance detection design to systematic scrutiny will enable social scientists to enhance their effective use in research.

Prompt engineering and in-context learning in LLMs

In-context learning, the ability of LLMs to generalize from a few examples provided within the prompt, has proven effective across a wide variety of language understanding tasks (Dong et al., 2024; J. Liu et al., 2022). Unlike conventional supervised machine learning, where models often require large amounts of annotated data for training, in-context learning allows models to adapt on the fly to new tasks by leveraging contextual information. This adaptability makes in-context learning particularly valuable for content annotation tasks, where efficient, low-cost solutions are often essential. However, the performance of prompt engineering and in-context learning depends on a combination of factors, such as the prompt template, the selection method, and the number of in-context examples, and the order in which these examples are presented (Zhao et al., 2021). In the present study, we evaluated the impact of different prompt templates (by varying the depth and detail of the instructions) and different shot selection strategies (by varying the number of in-context examples—zero-shot to many-shot—and whether a randomized and stratified shot sampling method was used to generate them) on model performance.

In prior studies, researchers have emphasized the importance of prompt template design, focusing on various features such as structure, specificity, clarity, and language framing within the instructions. Prompt templates may vary across dimensions like task framing, contextualization, and response style, each of which can affect the model's ability to capture task-specific language cues (Ma et al., 2022; Schick & Schütze, 2021a, 2021b). Building on these insights, we investigate whether using detailed prompt templates with high specificity and clear definitions for each level of stance detection will improve model performance across different LLMs. Specifically, the considerations of our prompt design alter four sub-dimensions to shift from a less detailed to a more detailed prompt template: (a) the role of the LLM, (b) classification definitions and details, (c) a wider range of language markers and forms, and (d) the importance of accurate classification.

Shot selection strategies also play a critical role in in-context learning performance: both the choice of examples to include in the prompt and the number of such examples, referred to as "shots," affect the model's understanding of the task. Zero-shot and few-shot learning has shown promise in content annotation tasks, with researchers seeking to improve the performance of these models through specific efforts at prompt engineering, often focusing on the prompt itself and the selection of shots used for in-context learning (Song et al., 2023). Others have focused on few-shot selection, devel-

oping strategies for appropriate example selection as another key element for in-context learning (An et al., 2023; Yao et al., 2023). Refining the selection of shots, both the number needed to provide sufficient context to the LLMs and the sampling strategy used to select these shots, has drawn research attention. While advanced machine learning methods for selecting in-context examples exist, such as k-NN-based unsupervised retrieval (J. Liu et al., 2022), they can be computationally demanding and costly to implement. To address this, our study utilizes two straightforward shot selection methods: random sampling and stratified sampling from the annotated training data. As widely applied by prior studies (Lu et al., 2022; Yao et al., 2023), random sampling provides a baseline approach by selecting examples directly from the social media dataset without any systematic/stratified strategy; while stratified sampling ensures that selected examples are stratified based on the stance levels within the dataset. By comparing these two approaches, we aim to understand the extent to which shot selection impacts the model's performance in detecting stances regarding HPV vaccination on social media.

The number of shots, or in-context examples, is another consideration in our study. Agarwal et al. (2024) found that increasing the number of shots (i.e., from few-shot to many-shot) yielded significant improvements in model performance across a wide variety of tasks, with more examples allowing the model to override pretraining biases and improve task specificity, asserting that "unlike few-shot learning, many-shot learning is effective at overriding pretraining biases" (p. 1). However, we believe that balancing the benefits of additional examples with computational efficiency remains a critical consideration, especially as LLMs develop through improvements in their underlying architecture and capabilities. Our study explores both few-shot and many-shot, as well as zero-shot, as baseline configurations to understand how shot quantity influences the model performance of the stance detection task.

In addition, we account for model sensitivity by testing several widely used state-of-the-art LLMs, both proprietary and open-source. Previous studies indicate that different LLMs exhibit varying sensitivity to in-context learning across tasks, such as sentiment analysis (W. Zhang et al., 2024). This variation can be attributed to differences in model architecture, pretraining data, and parameter count, which affect how models interpret and respond to prompt structures. Therefore, by evaluating multiple LLMs, our study aims to identify effective practices in prompt engineering and in-context learning that can be adapted across models.

Fine-tuning approaches in LLMs

Recognizing that in-context learning is not the only approach to enhancing the performance of LLMs for natural language processing and stance detection, our study also considers fine-tuning approaches for model comparison. Indeed, as a relatively new strategy, in-context learning was first used to describe the emergent behavior of LLMs only several years ago, following the rapid dilation of LLM data and model sizes (Brown et al., 2020; Dong et al., 2024). The more conventional and established strategy for adapting LLMs to correspond to a specific downstream task is fine-tuning. Fine-tuning is a process through which changes are introduced to some of the parameters of an LLM while holding the rest constant. By changing only a small fraction of the total model weights, an LLM receives new, and often more specialized, knowledge while maintaining the broad understanding of the model developed during pre-training, which remains accessible to the model through the unaltered weights. Fine-tuning allows a generalist LLM to hone its capabilities to complete a specified task, such as sentiment analysis (Dong et al., 2024; Mosbach et al., 2023).

Whether fine-tuning or in-context learning offers greater benefits to modern LLMs remains highly debated. Although in-context learning is often more computationally demanding during inference, it is unclear whether either approach consistently outperforms the other across tasks. Previous studies have shown that fine-tuned models outperform LLMs employing in-context learning on domain-specific tasks, while others have contradicted these findings (H. Liu et al., 2022; Mosbach et al., 2023; Yin et al., 2024). Furthermore, the distinction between the two approaches often depends on various factors, including model architecture, dataset quality, and the specific task assigned to the model (Mosbach et al., 2023; Tekumalla & Banda, 2023). However, few previous studies specifically examined the relative efficacy of fine-tuning versus in-context learning for stance detection of health information, as we do for social media discourse around HPV vaccination. Our study aims to systematically examine and compare the relative performance of fine-tuned and in-context learning-based models.

Additionally, given that a fine-tuned model's weights are altered for a specific task, contextualization is not needed during inference, which may lead to increased scalability and computational efficiency (Mosbach et al., 2023; Xia et al., 2024). The efficiency of fine-tuning is heightened if the fine-tuning technique used to modify the pre-trained LLM falls within the family of Parameter-Efficient Fine-Tuning (PEFT) methodologies, which significantly reduce computing time during the fine-tuning process as compared

to classical fine-tuning techniques (Xu et al., 2023). By incorporating and comparing in-context learning and fine-tuning, our study assesses the influence of prompt design, shot selection strategies, and fine-tuning methods on the performance of LLMs for stance detection of health information. We aim to provide practical guidance on refining LLM-based text annotation tasks for health information, demonstrating the potential for achieving high accuracy and efficiency in classifying complex and even controversial social media data.

Methods

Experimental design for prompt engineering and in-context learning

Our experimental design primarily tests three key dimensions of prompt design and in-context learning: prompt template complexity, shot sampling method, and shot quantity provided in the prompt. These dimensions are assessed across four widely used LLMs and their variants (i.e., GPT-4, Mistral, Llama 3, and Flan-UL2), chosen to capture variation in both model scale and architectural design, thereby allowing us to evaluate how different models adapt to in-context learning dimensions. In selecting LLMs, we considered both proprietary and open-source models to balance state-of-the-art performance, practical accessibility for the research community, and the ethics of data management. We also evaluated a range of model sizes to understand performance scaling and provide guidance to researchers working in resource-constrained settings. For instance, smaller open-source variants may be effectively deployed under computational constraints without a substantial loss of accuracy (Bai et al., 2024; Hsieh et al., 2023). Additionally, Flan-UL2 (20B), an instruction-tuned model trained on more than 1,800 diverse Natural Language Processing tasks (including text understanding, summarization, and generation), offers particularly strong zero-shot capabilities (W. Zhang et al., 2024). Based on these criteria, our final study examined the following four models, with two versions of GPT-4, Mistral, and Llama 3 being tested alongside Flan-UL2:

- (a) GPT-4: Turbo (gpt-4-0125-preview) and gpt-4o-mini;
- (b) Mistral: Mixtral-8x7B-Instruct-v0.1 and Mistral-7B-Instruct-v0.2;
- (c) Llama 3: Meta-Llama-3-70B-Instruct and Meta-Llama-3-8B-Instruct;
- (d) Flan-UL2.

Prompt template complexity

Prompt template complexity is defined as the level of detail of the prompt, with two levels tested: the basic prompt and the detailed prompt. In the basic prompt, we provide basic and essential information about HPV vaccination and instruct the model to classify the stance of a tweet regarding HPV vaccination. This prompt specifies that the stance should fall into one of the three categories: "in favor," "against," or "neutral or unclear." The detailed prompt builds on this by adding more guidance and information. In this version, we prompt the model to take on the role of "an expert content analyst," providing more specific and detailed definitions for each stance category, contextual information about HPV vaccination, and specifying the breadth of claims to consider ("statements, facts, statistics, opinions, or anecdotes"). This prompt also includes a caution to the model about the potential consequences of misclassifications, reinforcing the importance of accuracy. Examples of both prompt templates are listed in Appendix A.

Shot sampling method

The shot sampling method tested includes two approaches: random selection and stratified selection of examples. In random selection, a specified number of tweets is randomly drawn from the annotated training dataset, with no regard for stance balance. This approach, at large volumes, should better represent the distribution of content into the stance or topic categories. Stratified selection, on the other hand, involves randomly selecting a certain number of tweets representing each stance level ("in favor," "against," and "neutral or unclear") from the annotated training dataset to ensure a balanced representation of stance categories within the sample. This approach aims to expose the model to a more representative array of linguistic cues for each stance, which may improve classification accuracy.

Shot quantity

To evaluate the influence of shot quantity, we experiment with a range of shot counts, from 0 (zero-shot) up to 30 shots, at intervals of 3. In other words, the shot quantities tested include 0, 3, 6, 9, 12, 15, 18, 21, 24, 27, and 30. This systematic variation allows us to examine the trade-off between the depth of contextual examples and the classification performance of the model. By including zero-shot learning, we were able to assess the model's ability to infer stance without any supporting examples, providing a baseline to compare with few-shot and many-shot configurations.

Ground-truth data

The raw dataset used in the present study is Twitter data collected from Synthesio, a social listening platform, using a list of search terms (see Appendix B) related to HPV vaccination from January 1, 2023, to June 28, 2023. After collecting raw Twitter data (N = 313,900), three well-trained research assistants provided human annotations of a random sample (n = 1,050) for stance detection ("in favor," "against," or "neutral or unclear") toward HPV vaccination (intercoder reliability: Krippendorff's alpha = 0.70) as the ground-truth dataset for the study. We acknowledge that some tweet content may contain ambivalent and contradictory stances and information. To reduce stance ambivalence, we excluded tweets for which there was disagreement among the three research assistants on stance annotations. By doing this, the total size of this final ground-truth dataset was 756, including 367 tweets annotated as "in favor," 327 tweets annotated as "against," and 62 tweets annotated as "neutral or unclear" by the research assistants. The three annotators reached unanimous agreement on the annotations.

Procedures

We investigated the performance of various instruction fine-tuned LLMs in classifying the stance of the tweets toward HPV vaccination, using zero-shot and few-shot configurations, with varying shot quantities. The performance of the model within a given in-context learning scenario was evaluated using the macro F1 score, a metric that treats model performance across all three stance categories equally, regardless of class imbalance. Of all possible metrics, the macro F1 score aligns best with the aims of this study, as it weighs all three categories equally and is not influenced by the relative abundance of examples of each category in the evaluation dataset (Opitz, 2022).

In the zero-shot setting, LLMs were requested to classify any given tweet's stance without any examples provided, relying solely on the prompt instructions of the task, which included either a basic or detailed template. In contrast, the few-shot settings provided additional contextual guidance through a selected number of in-context examples, accompanying each task prompt. The presence of shots in the prompt allowed the LLMs to infer patterns in stance classification by observing in-context examples labeled as "in favor," "against," or "neutral or unclear" regarding HPV vaccination. To maintain data integrity and balance, we split the final annotated dataset into training and test sets in a stratified 50-50 manner. This ensured that stance categories were evenly represented across both sets, thus preserving

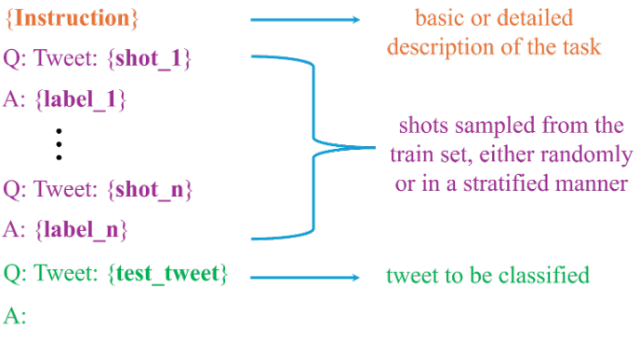

Figure 1: Overview of the prompt creation procedure.

stance balance for reliable performance evaluation. Tweets in the training set were used as few-shot examples within the prompts, while tweets in the test set were used to evaluate stance classification performance for each LLM.

To systematically evaluate the three dimensions in the experimental design, we created a comprehensive prompt dataset by embedding each test tweet within multiple prompt configurations. The total number of prompt configurations was determined by the combination of the following levels: prompt template complexity (basic vs. detailed), shot sampling methods (random vs. stratified), and shot quantity (ranging from 0 to 30 shots at intervals of 3). For each tweet in the test set, we generated 40 few-shot prompts and two zero-shot prompts, resulting in a dataset of 15,876 unique prompts. Specifically, each few-shot prompt for a test tweet was created by sampling examples from the training set independently. This approach ensured that every few-shot prompt provided unique combinations of contextual examples, maintaining diversity and reducing potential overlap that could lead to repetitive cues in classification. Figure 1 shows the high-level process for creating prompts for any test tweet.

Inference

Given that GPT-4 models are closed-source, we used OpenAI's APIs to send our inference requests directly to their servers. For all the other open-source models, Mistral, Llama 3, and Flan-UL2, we obtained their pre-trained weights from Hugging Face's Transformers library and conducted inference locally on the authors' server housed at their institute. This server was equipped with two NVIDIA RTX 6000 Ada Generation GPUs, each with

48GB of memory, enabling efficient handling of large-scale computations. For each LLM, we tokenized prompts using the respective tokenizer classes from Hugging Face's Transformers library to ensure compatibility with each LLM's specific architecture. Since these models support different maximum context lengths, we excluded any prompts that exceeded a model's input capacity. The maximal context lengths are as follows: 128,000 tokens for GPT-4 models, 8,192 tokens for Llama 3 models, 32,768 tokens for Mistral models, and 2,048 tokens for Flan-UL2.

Given that we aimed to optimize model focus and consistency for the classification task, we set the temperature parameter to zero or close to zero to minimize the randomness of the output. Specifically, we set the temperature to 0 for GPT-4 models, and 1e-5 for Llama 3, Mistral, and Flan-UL2 models. Additionally, we standardized certain parameters across these LLMs: 1) the batch size was set to 1 for inference processing speed; and 2) the maximum output length was capped at 200 tokens for all the models to control output verbosity. For memory efficiency, Mistral and Llama 3 models were loaded in 4 bits, reducing GPU memory consumption while maintaining computational efficiency.

Post-processing the LLM outputs

Since the LLM outputs do not always align exactly with the predefined stance labels, we employed a post-processing strategy using a pattern-matching tool to reliably extract the predicted stance label from each model output. Specifically, for each raw output, if the completion explicitly begins with or includes only one stance label, we treated the label as the model's prediction. Otherwise, in cases where multiple stance labels appeared in the response or where the response was ambiguous, we manually inspected the raw output and assigned the correct label. This human-in-the-loop approach ensured that the majority of completions were correctly categorized, allowing us to handle exceptions effectively and maintain high accuracy in the task of stance detection.

Fine-tuning methods and model selection

In the fine-tuning modeling, we employed a popular Parameter-Efficient Fine-Tuning (PEFT) technique called Low-Rank Adaptation (LoRA), which decomposes matrices within certain subunits of LLMs, reducing the number of trainable parameters in the model and the computational resources needed to fine-tune it (Hu et al., 2021). LLMs fine-tuned with LoRA perform as well as or better than models fine-tuned classically without using PEFT

techniques (Hu et al., 2022). Three models were chosen for fine-tuning with LoRA. These three models were chosen from amongst those used for in-context learning-based stance detection to serve as a representative sample of the models. The three models are the most performant models of all families of LLMs evaluated during the in-context learning-based stance analysis, with the exclusion of the GPT-4 Turbo, which is an OpenAI proprietary model and, therefore, cannot be fine-tuned. Hence, the fine-tuned models were Flan-UL2, Meta-Llama-3-70B-Instruct, and Mixtral-8x7B-Instruct-v0.1.

Additional analysis: In-context learning with subtopic-specific sampling

Building on prior literature that articulates the range of prevalent subtopics discussing HPV vaccination on social media, we examined whether specifying subtopics in both prompt and few-shot examples could improve model performance. For this analysis, we prepared a new set of prompts, varying both by subtopic (i.e., the three most prevalent in our dataset: reproductive health, cancer, and death) and by prompt template complexity (i.e., basic vs detailed). Given that each subtopic appeared in only a subset of the overall dataset, we expanded the ground truth beyond the unanimous-agreement dataset used in the main analysis to include additional cases where annotators did not reach full agreement. For these non-unanimous cases, we applied a majority-vote rule to determine the final label, thereby increasing data coverage and reducing potential bias toward training on less ambiguous tweets. To simplify class distribution and address imbalance, this additional analysis only focused on two stance levels, "in favor" and "against," while excluding the "unclear or neutral" cases. Full details of prompt design are provided in Appendix C, and case distributions for the few-shot pools and test sets in the subtopic-specific analysis are reported in Appendix D.

Results

In-context learning model performance

Figures 2A–G display the F1 scores for each LLM, plotted against the number of shots and prompt conditions (prompt template complexity and shot sampling methods) in in-context learning. Our results show varied performance trends across different models and conditions. Overall, GPT-4 Turbo presents the highest performance among all LLMs in our case. Notably, most models achieved acceptable F1 scores in the zero-shot learning conditions, though Mixtral-8x7B-Instruct-v0.1 falls short, with F1 scores below 0.70.

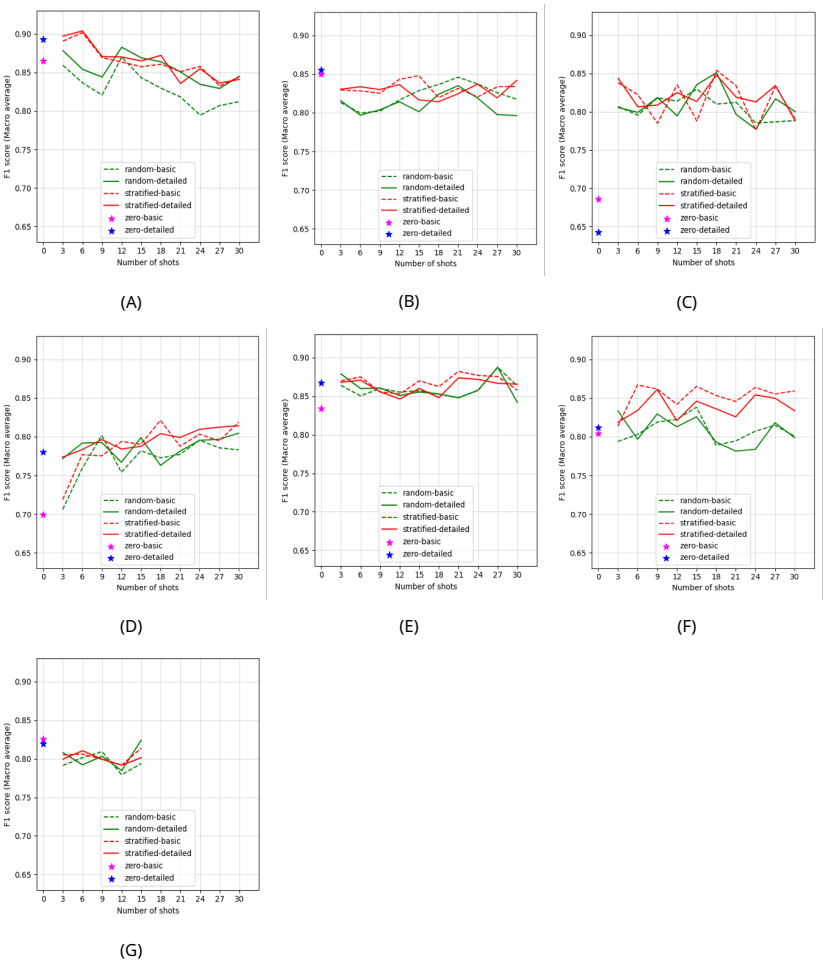

Figure 2: Performance of LLMs across varying experimental dimensions of in-context learning. (A) F1 scores of *GPT-4 Turbo* of in-context learning. (B) F1 scores of *GPT-4o-mini* of in-context learning. (C) F1 scores of *Mixtral-8x7B-Instruct* of in-context learning. (D) F1 scores of *Mistral-7B-Instruct* of in-context learning. (E) F1 scores of *Llama-3-70B-Instruct* of in-context learning. (F) F1 scores of *Llama-3-8B-Instruct* of in-context learning. (G) F1 scores of *Flan-UL2* of in-context learning.

However, performance improves across models by changing the number of shots, as well as varying the shot selection methods, though the extent of this improvement varies by model. Details of the model performance metrics are provided in Appendix G. In general, larger model variants perform better under the detailed-prompt condition (represented by solid lines in Figure 2), often outperforming the basic-prompt condition (dashed lines). In addition, the stratified sampling condition often yields higher F1 scores than random sampling, with the performance gap becoming more pronounced in smaller model variants.

Specifically, a unique trend appears with GPT-4 Turbo in Figure 2A, where F1 scores generally decline as the number of shots increases, except under the stratified sampling condition. Here, performance peaks at 6 shots, achieving the highest F1 score of 0.90 under the detailed prompt. Interestingly, increasing the number of shots does not improve model performance for GPT-4o-mini in Figure 2B, with the zero-shot condition performing the best (F1 scores = 0.85 and 0.86 in the basic and detailed prompts).

Zero-shot performance for Mixtral-8x7B-Instruct-v0.1 is comparatively lower, with F1 scores of 0.64 for a detailed prompt and 0.69 for a basic prompt, as shown in Figure 2C. However, by simply adding as few as three few-shot examples, F1 scores substantially improved by 0.20 points for detailed prompts, and by 0.12 for basic prompts. This result suggests that Mixtral-8x7B-Instruct-v0.1 is highly responsive to the presence of few-shot examples. We also tested its smaller variant, Mistral-7B-Instruct, as shown in Figure 2D, revealing that Mistral-7B-Instruct is particularly sensitive to prompt template complexity. This sensitivity results in a performance gap between basic and detailed prompts, especially in the zero- and 3-shot conditions.

In the case of Llama-3-70B-Instruct, as shown in Figure 2E, the overall performance pattern follows a "U-shaped curve," peaking at the 27-shot condition with stratified sampling. Few-shot configurations with fewer than 12 or more than 21 examples outperform those in the 15–18 range, along with a slight dip at 30 shots. Nevertheless, the zero-shot condition with a basic prompt underperforms other conditions in Llama-3-70B-Instruct. Additionally, Figure 2F shows the results of its smaller variant, Llama-3-8B-Instruct, with a pronounced sensitivity to shot sampling methods: stratified sampling conditions often outperform random sampling. This may suggest that stratified examples in the prompt help Llama-3-8B-Instruct acquire more balanced contextual knowledge across stance levels. Due to Flan-UL2's context length limitation of 2,048 tokens, we did not show the F1 scores for prompt configurations with more than 15 shots, as fewer than 100 tweets

met this context length criterion with higher shot counts. Nonetheless, we observe an uptick in performance for the detailed prompt at 15 shots, approaching zero-shot performance, though it still lags behind GPT-4 or Llama-3-70B-Instruct.

Additionally, the subtopic-specific design outperformed the above subtopic-agnostic design overall. For a fair comparison, we recomputed the F1-macro scores for only the "in favor" and "against" classes from the previous subtopic-agnostic results. Specifically, Mistral-7B-Instruct and Llama-3-8B-Instruct showed modest improvements with stratified sampling compared to random sampling, whereas their larger counterparts, Mixtral-8x7B-Instruct and Llama-3-70B-Instruct, did not show noticeable differences. Notably, we found that zero-shot performance in the Mistral models had a significant improvement under the subtopic-specific design, relative to the subtopic-agnostic baseline in the main analysis. For example, the F1 scores of Mixtral-8x7B-Instruct were: Subtopic-agnostic = 0.825; Subtopic-specific = 0.975 (reproductive health), 0.975 (cancer), and 1.000 (death). The F1 scores of Mistral-7B-Instruct were: Subtopic-agnostic = 0.850, Subtopic-specific = 0.925 (reproductive health), 0.925 (cancer), and 0.975 (death). Overall, across most conditions, subtopic-specific F1 scores were approximately 0.30 points higher than their subtopic-agnostic counterparts. More details are reported in Appendix E. Model performance metrics by class are provided in Appendix H.

During the procedure, we observed a small portion of ill-formatted outputs, such as "missing initial labels" and "dual stances," which can impact classification performance (de Wynter et al., 2023). The "missing initial label" outputs were most common in the zero-shot condition, where completions sometimes included unnecessary reasoning or failed to begin with a stance label. Overall, these ill-formatted cases accounted for approximately 1% of the outputs, and all were resolved through a human-in-the-loop process in which labels were manually corrected. In addition, in the subtopic-agnostic analysis, we found that explicitly instructing LLMs not to provide any extra texts reduced the ill-formatted outputs where more than one stance label was generated. After applying this human-in-the-loop process with manual inspection, we resolved all ill-formatted cases, ensuring high-quality stance labels.

Fine-tuning model comparison

After fine-tuning three models (i.e., Mixtral-8x7B-Instruct, Llama-3-70B-Instruct, and Flan-UL2) using LoRA, we found that Flan-UL2 achieved the

best performance based on macro F1 scores, narrowly outperforming the Mixtral model and surpassing Llama-3-70B-Instruct by a sizeable margin. The complete performance metrics of all three fine-tuned models are shown in Table 1. The superior performance of Flan-UL2 may seem initially surprising, as it performs markedly worse than the other two models on widely used LLM benchmarks (Grattafiori et al., 2024; A. Q. Jiang et al., 2024; Tay et al., 2022). However, given that more of Flan-UL2's parameters are accessible and amenable to LoRA adaptation compared to Mixtral-8x7B-Instruct and Llama-3-70B-Instruct, it is plausible that fine-tuning had a greater impact on its architecture, thereby boosting its task-specific performance. Additionally, model size may also play an important role. Both Flan-UL2 and Mixtral-8x7B-Instruct are significantly smaller than Llama-3-70B-Instruct in terms of parameters: Flan-UL2 and Mixtral-8x7B-Instruct have 20B and 47B parameters, respectively, while Llama-3-70B-Instruct has 70B parameters. We speculate that the smaller models, particularly Flan-UL2, adapted more effectively to fine-tuning on a relatively small and imbalanced stance classification dataset, while the much larger Llama-3-70B-Instruct may have required considerably more data to achieve similar gains (Kalajdzievski, 2024; Luo et al., 2025).

Overall, each fine-tuned model's performance during inference on the evaluation dataset was generally poorer than that of their untuned counterparts employing in-context learning across almost all experimental conditions. Even in cases where a fine-tuned model performed better than its in-context learning counterpart, such as Flan-UL2 compared to its few-shot learning counterpart, the advantage was only slight and inconsistent across different numbers of in-context examples. Our findings are consistent with the expectation that modern models like Llama-3-70B-Instruct and Mixtral-8x7B-Instruct, which are already trained on prompts with in-context examples, are intrinsically more receptive to and effective with in-context learning approaches (Grattafiori et al., 2024; A. Q. Jiang et al., 2024). Additionally, previous studies have also found that, when applied to smaller datasets, fine-tuned LLMs generally underperform relative to those employing in-context learning, further supporting the validity of the results in our study (Bertsch et al., 2025). Our fine-tuning analysis also examined quantization (i.e., 4-bit and 8-bit) but did not observe substantial improvements with higher precision (details in Appendix F).

Model	F1-micro	F1-macro	F1-weighted
Mixtral-8x7B-Instruct-v0.1	0.8919	0.8056	0.9025
Flan-UL2	0.8829	0.8126	0.8909
Meta-Llama-3-70B-Instruct	0.8649	0.7691	0.8787

Table 1: Complete Performance Metrics of the Three Models Fine-Tuned with LoRA

Discussion and implications

Our study contributes to the application of computational social science by testing the feasibility and providing guidance on integrating LLMs into social media annotation and stance detection of health information—in this case, HPV vaccine-related tweets—with systematic evaluation across models and conditions. It provides a framework for the careful application of LLMs to stance detection with attention to model fine-tuning and emergent in-context learning methods, systematically varying three key dimensions of prompt design and shot selection: prompt template complexity, shot sampling method, and shot quantity provided in the prompt. These dimensions are assessed across four widely used LLMs and their variants (i.e., GPT-4, Mistral, Llama 3, and Flan-UL2). We reveal a complex story, with particular models and design features yielding superior results and performances, suggesting dimensions and conditions researchers may wish to emphasize when integrating LLMs for the annotation of health information.

In the context of HPV vaccine-related tweets, we found that almost all in-context learning configurations outperform their fine-tuning counterparts. Notably, increasing shot quantity does not consistently enhance performance across models, especially for GPT-4 Turbo. Larger model variants tend to perform best under the detailed-prompt condition, typically outperforming the basic-prompt condition. In addition, the stratified sampling condition often outperforms random sampling, with the performance gap more pronounced in smaller model variants. Across tests, we find GPT-4 Turbo outperforms other frontier LLMs on overall performance metrics, with the best results emerging when using stratified sampling of "few-shot" in-context learning combined with a detailed contextual prompt. These findings should guide future research on health information classification across different LLMs, topics, or datasets, providing both a framework for future testing as LLMs continue to evolve and as dimensions to consider when conducting stance detection of health information or other social media content.

Taking the most widely used, and arguably the most advanced available LLM at the time of this analysis, GPT-4 Turbo, we found that the most efficient approach to stance detection is to provide six stratified examples (two per stance level), as opposed to a random selection of shots, along with a detailed, contextual prompt that asked the model to take on the role of "an expert content analyst," provided more specific and detailed definitions for each stance category alongside information about HPV vaccination, and specified the breadth of claims to consider ("statements, facts, statistics, opinions, or anecdotes"), while also cautioning the model about the potential consequences of misclassifications, reinforcing accuracy salience. In addition, when specifying subtopics (i.e., reproductive health, cancer, and death) in both prompts and few-shot examples, stance detection performance improved substantially, even after expanding analysis to include non-unanimous annotations, which suggests robustness in more detailed classification.

As noted, for GPT-4 Turbo, we observed a pattern of decreasing performance as the number of shots increased, suggesting that adding more examples may not necessarily improve stance detection for the most advanced LLMs. This observation highlights the importance of carefully calibrating in-context learning strategies rather than assuming that additional examples will automatically benefit LLM performance. It also offers reassurance for researchers with limited resources: effective performance can still be achieved with a modest number of validated examples, as our empirical evidence shows. It is worth noting that this phenomenon only occurred in GPT-4 Turbo and not in other open-source models we evaluated, so we were not able to draw any general conclusions without further investigation across additional topics and data.

As this is an observational finding, we are not able to provide a definitive explanation for the difference in LLM performance. Our speculations in terms of cognitive overload and human-labeled noise may help guide future research directions. The first speculation concerns cognitive overload in in-context learning, where more information may overwhelm LLMs' capacity for processing contextual information effectively (Upadhayay et al., 2024). Second, when adding more human-labeled examples for a few-shot learning, we may simultaneously introduce additional noise into LLMs' input through the process. While human annotations produced through rigorous procedures and reliability checks often serve as the gold standard, the reality is that social media messages are highly ambiguous and subject to human bias among annotators. Therefore, adding more examples does not

necessarily increase the overall prompt quality. Expanding the number of examples may increase rather than decrease uncertainty. Similar phenomena have also been documented by prior studies, which likewise speculate that data redundancy and noise of in-context learning can hinder model performance (Tang et al., 2025; X. Zhang et al., 2025). Future research should further investigate the mechanisms underlying the relationship between the number of shots and model performance in stance detection and other social media annotation tasks across additional contexts.

Additionally, we took fine-tuning methods into consideration and conducted model comparison with in-context learning performance. Our results show that fine-tuning models did not consistently outperform their untuned counterparts employing in-context learning. The limited improvement in model performance with fine-tuning compared to in-context learning in our study aligns with expectations, as modern models like Llama-3-70B-Instruct and Mixtral-8x7B-Instruct are pre-trained to work effectively with prompts containing in-context examples, making them inherently responsive to in-context learning methods (Grattafiori et al., 2024; A. Q. Jiang et al., 2024). Moreover, prior research has shown that on smaller datasets, fine-tuned LLMs often underperform relative to those employing in-context learning, further validating our comparison results (Bertsch et al., 2025). Nevertheless, fine-tuning may still be preferable to in-context learning models in certain scenarios, as fine-tuned models require substantially fewer computational resources for inference, making them faster and more scalable (Mosbach et al., 2023). This benefit is particularly crucial when computational resources are limited, processing tasks are enormous, or reducing the length of model inference is critical.

Future designs may want to scrutinize more nuanced aspects of prompt design and in-context learning when applying LLMs to different tasks and contexts. For instance, our current design included two conditions for prompt template complexity: basic and detailed prompts. The overall results showed that detailed-prompting conditions yielded better performance. In the detailed condition, we provided four additional details: 1) prompting the model to work as "an expert content analyst," 2) including detailed definitions of the three levels of stance (i.e., "in favor," "against," and "neutral or unclear") with specific contextual information about HPV vaccination, 3) specifying the breadth of claims to consider in stance detection ("statements, facts, statistics, opinions, or anecdotes"), and 4) alerting the model to costly consequences if misclassified. Since multiple elements were altered between the detailed-prompt condition and the basic-prompt condition,

it is unclear which aspect of the message design helped the most in terms of improving model performance. Future research should test more precise research designs to understand which components were consequential for LLMs' classification tasks. In addition, other factors of prompt design could be systematically altered and tested, further simplified prompt designs that omit contextual information, the inclusion of explicit information on the stance distribution in test and training data, and the incorporation of advanced machine learning methods for selecting in-context examples.

Lastly, researchers have addressed important ethical and data privacy concerns about using GPT models or other proprietary LLMs for annotation tasks (Yan et al., 2024). In our case, we relied on publicly available social media data collected through the social media listening platform Synthesio, and thus do not believe our use raises direct data privacy issues. However, future research using private or sensitive data should carefully evaluate the appropriateness of using proprietary LLMs for annotation tasks and should consult with their institutional review boards to ensure proper practices. We also acknowledge that the present study, serving as a case study of social media annotation in the health information domain, is not intended to encourage that OpenAI's GPT models or other proprietary LLMs should be prioritized over open-source models. Rather, researchers should consider their data characteristics, available computational resources, and ethical considerations, alongside the practical guidance offered here in our study (e.g., prompt design, shot selection and number, subtopic specification, and fine-tuning approaches) in making final decisions in methodologies. In particular, future work should weigh the trade-offs between using locally hosted open-source models, which can enhance data security, privacy, and reproducibility but may require greater technical capacity, and proprietary models like OpenAI's ChatGPT, which generally provide stronger model performance but pose greater challenges for data handling and compliance.

Conclusion

This study draws its conclusions from an experimental evaluation of stance detection of health information—specifically, HPV vaccine-related tweets—using prompt engineering and in-context learning, and comparing results to a fine-tuning approach. We compared both proprietary and open-source LLMs and their variants, including GPT-4 (Turbo and GPT-4o-mini), Mistral (Mixtral-8x7B-Instruct and Mistral-7B-Instruct), Llama 3 (Llama-3-70B-Instruct and Llama-3-8B-Instruct), and Flan-UL2, to reflect state-of-the-art performance and practical accessibility for the research community. Our

experimental design mainly focused on three dimensions: prompt template complexity (basic vs. detailed), shot sampling method (random vs. stratified), and shot quantity provided in the prompt (0 to 30, at intervals of 3). We also considered more conventional fine-tuning methods and subtopic specification for model comparison. While the evidence we present is restricted to our case of stance detection of HPV vaccine-related content, we believe the conclusions of this study extend to other health communication classification tasks and social media annotation across other social contexts.

Echoing Ziems et al. (2024), integrating LLMs for computational social science research, we further provide practical guidance on prompt design with detailed dimensions for social media annotations of health information. Generally, our findings highlight the value of employing in-context learning configurations over their fine-tuning counterparts. We find that more in-context learning examples do not guarantee better performance for stance detection, particularly for the most advanced model used in this analysis, GPT-4 Turbo. If the findings from this case are generalizable beyond the context of HPV vaccination, the best practice observed involves using detailed contextual prompts, stratified shot selection, and a "few-shot" selection approach to guide how state-of-the-art LLMs perform stance classification. When specifying subtopic domains (i.e., reproductive health, cancer, and death), stance performance improved substantially, even after expanding the dataset to include non-unanimous annotations, adding confidence to these findings, and suggesting robustness in more detailed classification. Other LLMs, such as Mistral, Llama 3, and Flan-UL2, exhibited different patterns of performance across the three main dimensions of in-context learning in the experiment, suggesting the need for continued research comparing how open-source and proprietary models perform.

Funding

This work was funded by the National Science Foundation through Convergence Accelerator Track F (Agency Tracking Number: 2230692; Award Number: MSN266268) and the John S. and James L. Knight Foundation (Award Number: MSN231314).

Appendices

All appendices, analysis scripts, and processed datasets are publicly available via the Open Science Framework (OSF) at https://doi.org/10.17605/OSF.IO/AHQ9V.

References

Agarwal, R., Singh, A., Zhang, L., Bohnet, B., Rosias, L., Chan, S., Zhang, B., Anand, A., Abbas, Z., Nova, A., et al. (2024). Many-shot in-context learning. *Advances in Neural Information Processing Systems*, *37*, 76930–76966.

An, S., Zhou, B., Lin, Z., Fu, Q., Chen, B., Zheng, N., Chen, W., & Lou, J.-G. (2023). Skill-based few-shot selection for in-context learning. *arXiv preprint arXiv:2305.14210*.

Bai, G., Chai, Z., Ling, C., Wang, S., Lu, J., Zhang, N., Shi, T., Yu, Z., Zhu, M., Zhang, Y., et al. (2024). Beyond efficiency: A systematic survey of resource-efficient large language models. *arXiv preprint arXiv:2401.00625*.

Bertsch, A., Ivgi, M., Xiao, E., Alon, U., Berant, J., Gormley, M. R., & Neubig, G. (2025). In-context learning with long-context models: An in-depth exploration. *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 12119–12149.

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.

Bussink-Voorend, D., Hautvast, J. L., Vandeberg, L., Visser, O., & Hulscher, M. E. (2022). A systematic literature review to clarify the concept of vaccine hesitancy. *Nature Human Behaviour*, *6*(12), 1634–1648.

Chen, L., Ling, Q., Cao, T., & Han, K. (2020). Mislabeled, fragmented, and conspiracy-driven: A content analysis of the social media discourse about the hpv vaccine in china. *Asian Journal of Communication*, *30*(6), 450–469.

Chuang, Y.-S., Wu, Y., Gupta, D., Uppaal, R., Kumar, A., Sun, L., Sreedhar, M. N., Yang, S., Rogers, T. T., & Hu, J. (2023). Evolving domain adaptation of pretrained language models for text classification. *arXiv preprint arXiv:2311.09661*.

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., et al. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*(11), 688–701.

de Wynter, A., Wang, X., Sokolov, A., Gu, Q., & Chen, S.-Q. (2023). An evaluation on large language model outputs: Discourse and memorization. *Natural Language Processing Journal*, *4*, 100024.

Dhaliwal, D., & Mannion, C. (2020). Antivaccine messages on facebook: Preliminary audit. *JMIR public health and surveillance*, *6*(4), e18878.

Di Domenico, G., Nunan, D., & Pitardi, V. (2022). Marketplaces of misinformation: A study of how vaccine misinformation is legitimized on social media. *Journal of public policy & marketing*, *41*(4), 319–335.

Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., et al. (2024). A survey on in-context learning. *Proceedings of the 2024 conference on empirical methods in natural language processing*, 1107–1128.

Du, J., Xu, J., Song, H.-Y., & Tao, C. (2017). Leveraging machine learning-based approaches to assess human papillomavirus vaccination sentiment trends with twitter data. *BMC medical informatics and decision making*, *17*(Suppl 2), 69.

Dunn, A. G., Surian, D., Leask, J., Dey, A., Mandl, K. D., & Coiera, E. (2017). Mapping information exposure on social media to explain differences in hpv vaccine coverage in the united states. *Vaccine*, *35*(23), 3033–3040.

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*.

Hanley, H., & Durumeric, Z. (2023). Tata: Stance detection via topic-agnostic and topic-aware embeddings. *Proceedings of the 2023 conference on empirical methods in natural language processing*, 11280–11294.

Heseltine, M., & Clemm von Hohenberg, B. (2024). Large language models as a substitute for human experts in annotating political text. *Research & Politics*, *11*(1), 20531680241236239.

Himelboim, I., Xiao, X., Lee, D. K. L., Wang, M. Y., & Borah, P. (2020). A social networks approach to understanding vaccine conversations on twitter: Network clusters, sentiment, and certainty in hpv social networks. *Health communication*, *35*(5), 607–615.

Hsieh, C.-Y., Li, C.-L., Yeh, C.-K., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C.-Y., & Pfister, T. (2023). Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *Findings of the Association for Computational Linguistics: ACL 2023*, 8003–8017.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. (2022). Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations (ICLR)*.

Hwang, J., Su, M.-H., Jiang, X., Lian, R., Tveleneva, A., & Shah, D. (2022). Vaccine discourse during the onset of the covid-19 pandemic: Topical structure and source patterns informing efforts to combat vaccine hesitancy. *Plos one*, *17*(7), e0271394.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., Casas, D. d. l., Hanna, E. B., Bressand, F., et al. (2024). Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Jiang, S., Wang, P., Liu, P. L., Ngien, A., & Wu, X. (2023). Social media communication about hpv vaccine in china: A study using topic modeling and survey. *Health Communication*, *38*(5), 935–946.

Jiang, X., Su, M.-H., Hwang, J., Lian, R., Brauer, M., Kim, S., & Shah, D. (2021). Polarization over vaccination: Ideological differences in twitter expression about covid-19 vaccine favorability and specific hesitancy concerns. *Social Media+ Society*, *7*(3), 20563051211048413.

Kalajdzievski, D. (2024). Scaling laws for forgetting when fine-tuning large language models. *arXiv preprint arXiv:2401.05605*.

King, A. J., Dunbar, N. M., Margolin, D., Chunara, R., Tong, C., Jih-Vieira, L., Matsen, C. B., & Niederdeppe, J. (2023). Global prevalence and content of information about alcohol use as a cancer risk factor on twitter. *Preventive medicine*, *177*, 107728.

Kornides, M. L., Badlis, S., Head, K. J., Putt, M., Cappella, J., & Gonzalez-Hernadez, G. (2023). Exploring content of misinformation about hpv vaccine on twitter. *Journal of Behavioral Medicine*, *46*(1), 239–252.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., et al. (2009). Computational social science. *Science*, *323*(5915), 721–723.

Lazer, D. M., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H., et al. (2020). Computational social science: Obstacles and opportunities. *Science*, *369*(6507), 1060–1062.

Liu, H., Tam, D., Muqeeth, M., Mohta, J., Huang, T., Bansal, M., & Raffel, C. A. (2022). Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, *35*, 1950–1965.

Liu, J., Shen, D., Zhang, Y., Dolan, W. B., Carin, L., & Chen, W. (2022). What makes good in-context examples for gpt-3? *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd workshop on knowledge extraction and integration for deep learning architectures*, 100–114.

Lu, Y., Bartolo, M., Moore, A., Riedel, S., & Stenetorp, P. (2022). Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8086–8098.

Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., & Zhang, Y. (2025). An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*.

Ma, R., Zhou, X., Gui, T., Tan, Y., Li, L., Zhang, Q., & Huang, X.-J. (2022). Template-free prompt tuning for few-shot ner. *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies*, 5721–5732.

Massey, P. M., Kearney, M. D., Hauer, M. K., Selvan, P., Koku, E., & Leader, A. E. (2020). Dimensions of misinformation about the hpv vaccine on instagram: Content and network analysis of social media characteristics. *Journal of medical Internet research*, *22*(12), e21451.

Mosbach, M., Pimentel, T., Ravfogel, S., Klakow, D., & Elazar, Y. (2023). Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. *arXiv preprint arXiv:2305.16938*.

Nan, X., & Madden, K. (2012). Hpv vaccine information in the blogosphere: How positive and negative blogs influence vaccine-related risk perceptions, attitudes, and behavioral intentions. *Health communication*, *27*(8), 829–836.

Opitz, J. (2022). From bias and prevalence to macro f1, kappa, and mcc: A structured overview of metrics for multi-class evaluation. *Heidelberg University*.

Ortiz, R. R., Smith, A., & Coyne-Beasley, T. (2019). A systematic literature review to examine the potential for social media to impact hpv vaccine uptake and awareness, knowledge, and attitudes about hpv and hpv vaccination. *Human vaccines & immunotherapeutics*, *15*(7-8), 1465–1475.

Piedrahita-Valdés, H., Piedrahita-Castillo, D., Bermejo-Higuera, J., Guillem-Saiz, P., Bermejo-Higuera, J. R., Guillem-Saiz, J., Sicilia-Montalvo, J. A., & Machío-Regidor, F. (2021). Vaccine hesitancy on social media: Sentiment analysis from june 2011 to april 2019. *Vaccines*, *9*(1), 28.

Puri, N., Coomes, E. A., Haghbayan, H., & Gunaratne, K. (2020). Social media and vaccine hesitancy: New updates for the era of covid-19 and globalized infectious diseases. *Human vaccines & immunotherapeutics*, *16*(11), 2586–2593.

Saulsberry, L., Fowler, E. F., Nagler, R. H., & Gollust, S. E. (2019). Perceptions of politicization and hpv vaccine policy support. *Vaccine*, *37*(35), 5121–5128.

Schick, T., & Schütze, H. (2021a). Exploiting cloze-questions for few-shot text classification and natural language inference. *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: main volume*, 255–269.

Schick, T., & Schütze, H. (2021b). It's not just size that matters: Small language models are also few-shot learners. *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies*, 2339–2352.

Semino, E., Coltman-Patel, T., Dance, W., Demjén, Z., & Hardaker, C. (2023). Pro-vaccination personal narratives in response to online hesitancy about the hpv vaccine: The challenge of tellability. *Discourse & Society*, *34*(6), 752–771.

Smith, T. C., & Gorski, D. H. (2024). Infertility: A common target of antivaccine misinformation campaigns. *Vaccine*, *42*(4), 924–929.

Song, Y., Wang, T., Cai, P., Mondal, S. K., & Sahoo, J. P. (2023). A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Computing Surveys*, *55*(13s), 1–40.

Sun, L., Chuang, Y.-S., Sun, Y., & Yang, S. (2023). Adoption and implication of the biased-annotator competence estimation (bace) model into covid-19 vaccine twitter data: Human annotation for latent message features. *arXiv preprint arXiv:2302.09482*.

Surian, D., Nguyen, D. Q., Kennedy, G., Johnson, M., Coiera, E., & Dunn, A. G. (2016). Characterizing twitter discussions about hpv vaccines using topic modeling and community detection. *Journal of medical Internet research*, *18*(8), e232.

Tan, Z., Li, D., Wang, S., Beigi, A., Jiang, B., Bhattacharjee, A., Karami, M., Li, J., Cheng, L., & Liu, H. (2024). Large language models for data annotation and synthesis: A survey. *arXiv preprint arXiv:2402.13446*.

Tang, Y., Tuncel, D., Koerner, C., & Runkler, T. (2025). The few-shot dilemma: Over-prompting large language models. *arXiv preprint arXiv:2509.13196*.

Tay, Y., Dehghani, M., Tran, V. Q., Garcia, X., Wei, J., Wang, X., Chung, H. W., Shakeri, S., Bahri, D., Schuster, T., et al. (2022). Ul2: Unifying language learning paradigms. *arXiv preprint arXiv:2205.05131*.

Tekumalla, R., & Banda, J. M. (2023). Leveraging large language models and weak supervision for social media data annotation: An evaluation using covid-19 self-reported vaccination tweets. *International Conference on Human-Computer Interaction*, 356–366.

Törnberg, P. (2025). Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*, *43*(6), 1181–1195.

Upadhayay, B., Behzadan, V., & Karbasi, A. (2024). Cognitive overload attack: Prompt injection for long context. *arXiv preprint arXiv:2410.11272*.

Vraga, E. K., Brady, S. S., Gansen, C., Khan, E. M., Bennis, S. L., Nones, M., Tang, R., Srivastava, J., & Kulasingam, S. (2023). A review of hpv and hbv vaccine hesitancy, intention, and uptake in the era of social media and covid-19. *eLife*, *12*, e85743.

Wang, X., Chen, L., Shi, J., & Peng, T.-Q. (2019). What makes cancer information viral on social media? *Computers in Human Behavior*, *93*, 149–156.

Weinzierl, M., & Harabagiu, S. (2022). Vaccinelies: A natural language resource for learning to recognize misinformation about the covid-19 and hpv vaccines. *arXiv preprint arXiv:2202.09449*.

Xia, Y., Kim, J., Chen, Y., Ye, H., Kundu, S., Hao, C. C., & Talati, N. (2024). Understanding the performance and estimating the cost of llm fine-tuning. *2024 IEEE International Symposium on Workload Characterization (IISWC)*, 210–223.

Xu, L., Xie, H., Qin, S.-Z. J., Tao, X., & Wang, F. L. (2023). Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *arXiv preprint arXiv:2312.12148*.

Yan, B., Li, K., Xu, M., Dong, Y., Zhang, Y., Ren, Z., & Cheng, X. (2024). On protecting the data privacy of large language models (llms): A survey. *arXiv preprint arXiv:2403.05156*.

Yao, B., Chen, G., Zou, R., Lu, Y., Li, J., Zhang, S., Sang, Y., Liu, S., Hendler, J., & Wang, D. (2023). More samples or more prompts? exploring effective in-context sampling for llm few-shot prompt engineering. *arXiv preprint arXiv:2311.09782*.

Yin, Q., He, X., Leong, C. T., Wang, F., Yan, Y., Shen, X., & Zhang, Q. (2024). Deeper insights without updates: The power of in-context learning over fine-tuning. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 4138–4151.

Zhang, H., Wheldon, C., Dunn, A. G., Tao, C., Huo, J., Zhang, R., Prosperi, M., Guo, Y., & Bian, J. (2020). Mining twitter to assess the determinants of health behavior toward human papillomavirus vaccination in the united states. *Journal of the American Medical Informatics Association*, *27*(2), 225–235.

Zhang, W., Deng, Y., Liu, B., Pan, S., & Bing, L. (2024). Sentiment analysis in the era of large language models: A reality check. *Findings of the Association for Computational Linguistics: NAACL 2024*, 3881–3906.

Zhang, X., Lv, A., Liu, Y., Sung, F., Liu, W., Luan, J., Shang, S., Chen, X., & Yan, R. (2025). More is not always better? enhancing many-shot in-context learning with differentiated and reweighting objectives. *arXiv preprint arXiv:2501.04070*.

Zhao, Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models. *International conference on machine learning*, 12697–12706.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, *50*(1), 237–291.