Stance detection enhanced by large language models

by

Nikesh Gyawali

B.E., Tribhuvan University, Nepal, 2016

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Computer Science
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2025

Abstract

Stance detection, a subfield of opinion mining, is an important task in Natural Language Processing (NLP) that involves determining the attitude or position expressed by an author of a text towards a particular topic or claim, generally referred to as target. More specifically, the task involves automatically determining if an author of a text is *In-Favor (Positive)*, *Against (Negative)*, or *Neutral* towards a given target. Accurate stance detection is crucial for understanding complex social issues, guiding policy decisions, and shaping effective interventions. While the field of NLP has seen significant progress, capturing nuanced stances from diverse and often ambiguous data remains a challenge. Recent developments in Large Language Models (LLMs) have transformed the field by offering unprecedented capabilities in interpreting subtle linguistic cues and context-dependent meanings. This dissertation advances the field of stance detection by utilizing Large Language Models (LLMs) to improve methodologies across multiple critical and socially impactful domains.

In this dissertation, we first explore LLM-enhanced approaches to stance detection in the socially controversial and polarizing topics of gun regulation and vaccines. We introduce the GunStance dataset, consisting of social media posts from X (formerly Twitter) posted by X users after major mass shooting events in the United States. By integrating labeled and unlabeled posts, this dataset allows comprehensive exploration of a semi-supervised learning framework in the context of stance detection. We propose a novel hybrid model that combines semi-supervised techniques with LLMs and show that our approach significantly outperforms traditional stance detection approaches.

Furthermore, we assemble a large dataset of social media posts from X, capturing the vaccine discourse over a decade. The dataset includes seven years before COVID-19, as well as three COVID-19 years. Leveraging LLMs, social cognition theories and emotional dynamics, we analyze the vaccine dataset to capture the evolving public attitudes towards vaccines

before and during the COVID-19 pandemic. Our study reveals increasing polarization and heightened emotional engagement, with a notable rise in vaccine skepticism amid the global health crisis.

Expanding into the less controversial but important financial domain, we construct a financial stance detection corpus from annual 10-K reports filed to U.S. Securities and Exchange Commission (SEC) and earnings call transcripts (ECT) by extracting short text fragments relevant to key financial metrics, such as debt, earnings per share (EPS), and sales, and annotating them using LLM-driven methodologies with strict human validation. This financial stance detection corpus facilitates extensive evaluation of LLMs' ability to detect subtle stances towards financial metrics, a task that requires complex reasoning. Our findings demonstrate the effectiveness of LLMs in performing accurate stance detection without extensive labeled data, showcasing their potential for real-world financial analysis applications.

Building upon these insights, we also introduce the Modular Prompt Optimization for Stance Detection (MoPrO-SD) framework. This framework utilizes the prompt optimization capabilities of LLMs by breaking down the complex stance detection prompt into modular, optimizable components. Each module is iteratively refined using LLMs as prompt optimizers, leading to an improved prompt that outperforms human-crafted prompts on several stance detection benchmarks.

Collectively, this dissertation advances the field of stance detection by providing comprehensive evidence on the use of LLMs to enhance the performance, adaptability, and efficiency of stance detection methodologies across social media posts and financial documents, offering an analytical and scalable framework for informed and nuanced decision-making in an increasingly digital and interconnected world.

Stance detection enhanced by large language models

by

Nikesh Gyawali

B.E., Tribhuvan University, Nepal, 2016

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Computer Science
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2025

Approved by:

Major Professor
Doina Caragea

Copyright

Abstract

Stance detection, a subfield of opinion mining, is an important task in Natural Language Processing (NLP) that involves determining the attitude or position expressed by an author of a text towards a particular topic or claim, generally referred to as target. More specifically, the task involves automatically determining if an author of a text is *In-Favor (Positive)*, *Against (Negative)*, or *Neutral* towards a given target. Accurate stance detection is crucial for understanding complex social issues, guiding policy decisions, and shaping effective interventions. While the field of NLP has seen significant progress, capturing nuanced stances from diverse and often ambiguous data remains a challenge. Recent developments in Large Language Models (LLMs) have transformed the field by offering unprecedented capabilities in interpreting subtle linguistic cues and context-dependent meanings. This dissertation advances the field of stance detection by utilizing Large Language Models (LLMs) to improve methodologies across multiple critical and socially impactful domains.

In this dissertation, we first explore LLM-enhanced approaches to stance detection in the socially controversial and polarizing topics of gun regulation and vaccines. We introduce the GunStance dataset, consisting of social media posts from X (formerly Twitter) posted by X users after major mass shooting events in the United States. By integrating labeled and unlabeled posts, this dataset allows comprehensive exploration of a semi-supervised learning framework in the context of stance detection. We propose a novel hybrid model that combines semi-supervised techniques with LLMs and show that our approach significantly outperforms traditional stance detection approaches.

Furthermore, we assemble a large dataset of social media posts from X, capturing the vaccine discourse over a decade. The dataset includes seven years before COVID-19, as well as three COVID-19 years. Leveraging LLMs, social cognition theories and emotional dynamics, we analyze the vaccine dataset to capture the evolving public attitudes towards vaccines

before and during the COVID-19 pandemic. Our study reveals increasing polarization and heightened emotional engagement, with a notable rise in vaccine skepticism amid the global health crisis.

Expanding into the less controversial but important financial domain, we construct a financial stance detection corpus from annual 10-K reports filed to U.S. Securities and Exchange Commission (SEC) and earnings call transcripts (ECT) by extracting short text fragments relevant to key financial metrics, such as debt, earnings per share (EPS), and sales, and annotating them using LLM-driven methodologies with strict human validation. This financial stance detection corpus facilitates extensive evaluation of LLMs' ability to detect subtle stances towards financial metrics, a task that requires complex reasoning. Our findings demonstrate the effectiveness of LLMs in performing accurate stance detection without extensive labeled data, showcasing their potential for real-world financial analysis applications.

Building upon these insights, we also introduce the Modular Prompt Optimization for Stance Detection (MoPrO-SD) framework. This framework utilizes the prompt optimization capabilities of LLMs by breaking down the complex stance detection prompt into modular, optimizable components. Each module is iteratively refined using LLMs as prompt optimizers, leading to an improved prompt that outperforms human-crafted prompts on several stance detection benchmarks.

Collectively, this dissertation advances the field of stance detection by providing comprehensive evidence on the use of LLMs to enhance the performance, adaptability, and efficiency of stance detection methodologies across social media posts and financial documents, offering an analytical and scalable framework for informed and nuanced decision-making in an increasingly digital and interconnected world.

Table of Contents

List of Figures xii

List of Tables xiv

Acknowledgements xvi

1 Introduction 1

1.1 Overview 1

1.2 Contributions of this dissertation 4

2 GunStance: Stance detection for gun control and gun regulation 9

2.1 Introduction 9

2.2 Related work 11

2.2.1 Stance detection datasets 11

2.2.2 Stance detection approaches 12

2.3 GunStance dataset 13

2.4 Methodology 19

2.4.1 Proposed approach: hybrid SSL/LLM 19

2.4.2 Baseline models 22

2.4.3 Experimental setup 22

2.5 Results and discussion 23

2.6 Conclusions and future work 25

3 The shifting landscape of vaccine discourse: A massive dataset examining a decade of English-language social media posts on vaccines before and during the COVID-19 pandemic . . . 28
3.1 Introduction . . . 28
3.2 Related work . . . 33
3.3 Methodology . . . 38
3.3.1 Utterance Emotion Dynamics (UED) metrics . . . 39
3.3.2 Emotion lexicons . . . 40
3.3.3 Stance detection . . . 41
3.4 Results and discussion . . . 42
3.5 Conclusions . . . 53
3.6 Limitations and future research . . . 53

4 Evaluating large language models for stance detection on financial targets from SEC filing reports and earnings call transcripts . . . 56
4.1 Introduction . . . 56
4.2 Related work . . . 58
4.3 Financial dataset . . . 59
4.3.1 SEC form 10-K annual report . . . 59
4.3.2 Quarterly earnings call transcripts . . . 60
4.3.3 Dataset collection . . . 61
4.4 Methodology . . . 62
4.4.1 Large language models . . . 62
4.4.2 Experimental setup . . . 64
4.5 Results and discussion . . . 66
4.5.1 Context usage: no context, full context, and summarized context . . . 66
4.5.2 Random vs. semantically similar few-shot examples . . . 68
4.5.3 Effectiveness of Chain-of-Thought (CoT) prompting . . . 69

4.5.4 Few-shot prompting vs. zero-shot prompting . . . 70
4.5.5 Performance across SEC and ECT datasets . . . 71
4.6 Conclusions . . . 72
4.7 Limitations . . . 73

5 Modular prompt optimization for stance detection . . . 74
5.1 Introduction . . . 74
5.2 Related work . . . 77
5.3 Methodology . . . 79
5.3.1 Stance detection prompt modules . . . 80
5.3.2 Proposed MoPrO-SD framework . . . 80
5.3.3 Methods of prompt optimization . . . 81
5.4 Experimental setup . . . 83
5.4.1 Stance detection datasets . . . 83
5.4.2 Large language models . . . 85
5.4.3 Research questions . . . 86
5.5 Results and discussion . . . 86
5.6 Conclusion . . . 91

6 Concluding remarks and future prospects . . . 93

Bibliography . . . 96

A Appendix: GunStance: Stance detection for gun control and gun regulation . . . 126
A.1 Existing stance detection datasets . . . 126
A.2 Crawling rules/terms . . . 126
A.3 Tweets statistics . . . 130
A.4 Baseline models . . . 130
A.4.1 Semi-supervised baselines . . . 132

A.4.2 Zero-shot baseline models . . . 133
A.5 Experimental setup . . . 134

B Appendix: The shifting landscape of vaccine discourse: A massive dataset examining a decade of social media posts on vaccines before and during the COVID-19 pandemic . . . 138
B.1 Prompt for Llama 3.3 model . . . 138
B.2 COVID-19 vaccine timeline . . . 141
B.3 Related work . . . 143
B.4 Top 15 low warmth and dominance words . . . 144

C Appendix: Evaluating Large Language Models for stance detection on financial targets from SEC filing reports and earnings call transcripts . . . 145
C.1 LLM prompt for relevance filtering . . . 145
C.2 LLM prompt for stance annotation . . . 147
C.3 Examples of dataset instances . . . 150
C.4 Chain-of-Thought example . . . 151
C.5 Error analysis . . . 153
C.5.1 Zero-shot vs. few-shot . . . 153
C.5.2 CoT vs. non-CoT . . . 154
C.5.3 Random vs. similar example usage . . . 154
C.6 All experiment results . . . 156

D Appendix: Modular prompt optimization for stance detection . . . 162
D.1 Meta prompts . . . 163
D.2 Complete prompt . . . 167

List of Figures

2.1 Bi-gram word clouds for "banning guns" 17

2.2 Stance-wise moral and sentiment scores 18

2.3 F1-scores in the cross-target setup 26

3.1 Monthly trends in Twitter vaccine posts and engagement 37

3.2 Distribution of word counts and post lengths. 38

3.3 Monthly trend in positive and negative sentiment words from 2013–2022. . . 43

3.4 Monthly trend in emotion-word density during the COVID-19 44

3.5 Comparison of vaccine-discourse "home bases" before and during COVID-19 47

3.6 Warmth and competence language trends in vaccine discourse 48

3.7 Monthly proportion of pro- and anti-vaccine posts from 2013–2022 51

4.1 Zero-shot accuracy on SEC and ECT datasets (with/without CoT prompting) 67

4.2 Context usage scenarios across different models on two datasets 68

4.3 Few-shot transcript usage scenario accuracy (with/without CoT prompts) . . 69

4.4 Few-shot with chain-of-thought accuracy on two datasets across various targets 70

5.1 Overview of the MoPrO-SD framework 79

5.2 Validation accuracy over 10 optimization steps on VaccineStance dataset . . 87

5.3 Zero- and few-shot prompting performance across four benchmark datasets . 88

A.2.1 Top 20 news outlets by tweet count 128

A.3.2 Distribution of the words per tweet and the total character length of the tweet 130

A1 Top 15 low-warmth words by vaccine stance (in-favor vs. against) 144

A2 Top 15 low-competence words by vaccine stance (in-favor vs. against) 144

A1 Example of a complete base prompt for stance detection 167

List of Tables

2.1 Annotated tweet examples for GunStance dataset 13

2.2 Labelled and unlabelled tweet counts for GunStance Dataset 16

2.3 Statistics of dataset splits for GunStance dataset 18

2.4 Performance comparison of baseline models for GunStance dataset 21

2.5 Performance comparison for the leave-one-event-out models 24

3.1 Emotion-word density of various emotions before and during COVID-19 . . . 45

3.2 Per-class and overall stance classification performance of Llama 3.3 49

3.3 Example posts labeled by Llama 3.3 and human annotators 50

3.4 Emotion-word density for low-warmth and low-competence emotions 52

4.1 Sample instances from ECT and SEC datasets 60

4.2 Distribution of stance categories for each financial target 61

4.3 Few-shot classification accuracy (%) on the ECT and SEC datasets 66

5.1 Example posts labeled by Llama 3.3 and human annotators 83

5.2 Results of prompting strategy on SemEval-2016 dataset 89

5.3 Results of prompting strategy on GunStance dataset 89

5.4 Results of prompting strategy on COVID-19-Stance dataset 90

5.5 Results of prompting strategy on Vaccines dataset 91

5.6 Accuracy of prompt optimization across datasets with two reasoning models 92

A1 Summary of public stance-detection datasets 127

A2 Seed rules for each shooting event 128

A3 Additional rule-based hashtags 129

A1 Summary of Related Works on Vaccines and COVID-19 post dataset 143

A1 ECT and SEC example instances 150

A1 Sample distribution across train, validation, and test sets for each target in the stance detection datasets. 162

Acknowledgments

I owe deep gratitude to the many people who made my PhD years so rewarding. First and foremost, I would like to thank my supervisor, Professor Doina Caragea, whose support and guidance have been immense throughout this journey. Through hundreds of emails and countless hours of meetings, she has helped shape my path from a curious beginner into the researcher I am today. With patience, she reviewed my early, often terrible drafts–sometimes more than once–and provided valuable feedback. She consistently pushed me to think critically, develop meaningful research ideas, and present my findings clearly and effectively in research papers.

I am also grateful to my PhD advisory committee members: Professor George Amariucai, Professor Cornelia Caragea, and Professor Eric Fitzsimmons for their valuable time, insights, and encouragement.

I have been fortunate to learn from many inspiring individuals who have supported me throughout my PhD journey. I had the privilege of working with Professor Sanzhen Liu on applying deep learning to bioinformatics over two summers, and with Professor Husain Aziz on applying machine learning to transportation safety. Likewise, Professors Eugene Vasserman, William Hsu, and Stephen Welch provided valuable guidance during my early PhD years.

I am thankful to my collaborators, with whom I have had the pleasure of working and learning from: Ravi Bika, Iustin Sirbu, Tiberiu Sosea, Sarthak Khanal, Traian Rebedea, Saif Mohammad, Aditya Jha, and Qais Tasali.

I have been fortunate to meet and be friends with many wonderful people along the way. I am especially grateful to my friends in Manhattan, KS, across the US, and back home in Nepal for their unwavering support and friendship.

I am grateful to funding agencies that supported my research, including the Federal Motor Carrier Safety Administration (FMCSA) and the National Science Foundation (NSF)

through STTR grant 2343777 and grants IIS-2107487, IIS-1741345, and ITE-2137846.

Finally, and most importantly, I owe my deepest gratitude to my brother, my parents, and my grandparents, who have always believed in me even more than I believed in myself. Their unwavering love, encouragement, and faith have been my anchor through every challenge and my inspiration in every success.

Chapter 1

Introduction

1.1 Overview

Stance detection, a subfield of opinion mining, is an important task in Natural Language Processing (NLP) that involves determining the attitude or position expressed by the author of a text towards a particular topic or claim, generally referred to as target. More specifically, the task involves automatically determining if an author of a text is *In-Favor (Positive)*, *Against (Negative)*, or *Neutral* towards a given target (Mohammad et al., 2016). Stance detection has broad applications, ranging from misinformation detection and public opinion mining on political and health issues (ALDayel and Magdy, 2021; Küçük and Can, 2020) to extracting the key performance of a company towards financial targets from financial documents (Du et al., 2024). The proliferation of digital platforms has yielded vast, heterogeneous textual data, making stance detection an essential analytical method with broad implications.

Historically, stance detection primarily relied on supervised machine learning approaches that involved manual feature engineering, lexicon-based methods, and utilization of extensive annotated datasets. While influential, these traditional methods face limitations, including difficulty in capturing implicit stance expressions, sarcasm, irony, and contextual nuances in human languages. Moreover, the necessity of large-scale annotated datasets makes these

techniques resource-intensive and inflexible, particularly when rapid analysis of emerging societal issues is necessary (ALDayel and Magdy, 2021; Mets et al., 2024). Such limitations have motivated the exploration of more advanced, data-scarce computational approaches.

The emergence of Large Language Models (LLMs) has transformed the field of NLP by offering unprecedented capabilities in understanding subtle linguistic cues and context-dependent meanings and addressing the limitations of traditional stance detection methods. LLMs are pre-trained on extensive and diverse textual data and demonstrate superior proficiency in semantic understanding, natural language interpretation, and reasoning capabilities. Furthermore, their strong performance in zero-shot and few-shot learning contexts significantly reduces the requirement of extensive labeled data (Pangtey et al., 2025; Wang and Brorsson, 2025), thus enhancing the adaptability and scalability of the stance detection task across varied domains.

Motivated by the potential of LLMs, in this dissertation, we systematically explore how LLMs can enhance stance detection by leveraging their semantic understanding, reasoning capabilities, and adaptability to novel contexts. This research bridges fundamental NLP advancements with pragmatic applications in impactful societal domains such as the polarizing social issues of gun control and regulation, human vaccines in relation to public health, and finance. We demonstrate the broader applicability and practical utility of LLM-based methodologies for stance detection. The remainder of this dissertation is organized in several chapters, each devoted to one study (i.e., publication).

In Chapter 2, we explore stance detection surrounding the controversial topics of gun control and regulation using posts on X (formerly Twitter) following major mass shooting incidents in the United States. Specifically, we introduce a novel `GunStance` dataset, comprising both labeled and unlabeled social media posts from X. This dataset enables the exploration of semi-supervised and hybrid methodologies that integrate traditional approaches with the strengths of LLMs. Empirical results demonstrate the ability of the proposed approach to provide significant improvement in detecting nuanced stance in the social posts, providing a practical, efficient, and scalable solution for monitoring public discourse on gun control and regulation.

In Chapter 3, we explore the vaccine-related discourse in social media that is particularly significant in connection to public health, as was recently underlined by the COVID-19 pandemic. Vaccine hesitancy and misinformation represent urgent societal challenges, making the accurate detection of public stances towards vaccines critical for health communication and policy formulation strategies. Using a robust longitudinal dataset containing 18.7 million social media posts spanning a decade, we show how social cognition theories and frameworks of emotional dynamics can effectively identify and help interpret complex shifts in public opinion towards vaccines. We further utilize LLM-driven methodologies to study how the stance towards vaccines changed over a decade, pre-COVID-19 (2013-2019) and during the COVID-19 pandemic (2020-2022). These insights reveal critical patterns in emotional polarization, evolving skepticism, and public trust towards vaccines.

Building upon insights from social media posts, we extend the methodological framework into the financial domain in Chapter 4. We focus on corporate narratives in financial disclosures such as SEC filings and earnings call transcripts (ECT). Financial texts are characterized by intricate jargon, domain-specific terminologies, and nuanced languages, posing unique challenges for accurate stance detection. This dissertation contributes a carefully annotated sentence-level corpus for specific financial targets–debt, earnings per share (EPS), and sales–and leverages the advanced capabilities of LLMs for systematic evaluation and analysis. Findings from this research show how LLM-enhanced stance detection can effectively decode subtle managerial attitudes and risk assessments, providing a valuable analytical tool for investors, auditors, and regulatory bodies.

Synthesizing insights from these diverse domains, in Chapter 5, we propose a novel Modular Prompt Optimization for Stance Detection (MoPrO-SD) framework tailored to utilize the reasoning and automatic prompt optimization capabilities of LLMs. This modular framework systematically decomposes a long prompt into distinct modules, optimizing each module through an iterative refinement process using meta-prompts and leveraging the LLM as a prompt optimizer. Comprehensive evaluations across various benchmarks and LLM models show the effectiveness of our prompt optimization framework, which consistently yields superior performance as compared to an approach that relies on human-crafted base

prompts.

Collectively, this dissertation advances the field of stance detection by systematically demonstrating how LLMs can significantly enhance the performance, adaptability, and efficiency of stance detection methodologies. By providing comprehensive evidence across social media posts with polarizing topics such as gun control and regulation, vaccine discourse, and financial disclosures, we highlight the capabilities of LLMs, offering scalable, analytical frameworks that support informed and nuanced decision-making in an increasingly digital and interconnected world.

1.2 Contributions of this dissertation

This dissertation unifies four empirical studies that explore and extend the stance detection research across supervised, semi-supervised, and LLM-based paradigms. The key contributions of this dissertation are outlined below:

Methodological contributions

1. **Hybrid semi-supervised/LLM approach**: We introduce a hybrid model that combines traditional semi-supervised learning methods with advanced LLM capabilities. This approach utilizes the reasoning and contextual understanding capability of LLMs to selectively generate pseudo-labels for unlabeled data. Experimental results show that the proposed hybrid semi-supervised/LLM approach significantly improves stance detection performance.

2. **Modular prompt optimization for stance detection (MoPrO-SD)**: We introduce a novel methodological framework that enables modular prompt optimization using LLMs. MoPrO-SD systematically breaks complex stance detection prompts into smaller, optimizable modules, each of which is optimized iteratively using LLMs. This approach significantly improves model performance and adaptability across various domains.

Resource and dataset contributions

1. **GunStance dataset**: We curate a comprehensive dataset comprising labeled and unlabeled social media posts from X (formerly Twitter) related to gun control and regulation debates following various mass shooting events in the United States. This dataset serves as a valuable resource for exploring semi-supervised stance detection methodologies in the polarizing social topic of gun control and regulation.

2. **Vaccine posts and stance detection datasets**: We curate an extensive dataset containing 18.7 million social media posts from X (formerly Twitter) spanning over a decade. The dataset is rigorously preprocessed to facilitate a robust longitudinal study of public vaccine discourse. We also construct a vaccine stance detection dataset by manually annotating a subset of 2000 posts (using paid crowd-sourcing services) in terms of stance towards the vaccine target.

3. **Financial stance detection dataset**: We construct a specialized sentence-level corpus derived from SEC filing reports and earnings call transcripts for stance detection towards key financial metrics (specifically, debt, earnings per share, and sales). This dataset provides a valuable resource for advancing research in fine-grained financial text analysis and the development of robust stance detection models in the financial domain.

Empirical insights and practical applications

1. **Polarized gun-related public debates**: Our study on LLM-driven methodologies in accurately detecting stances in highly contentious social debates regarding gun control and regulation can offer practical benefits for policymakers, advocacy groups, and researchers in understanding the polarization in social media.

2. **Public health communication**: We perform an in-depth analysis of vaccine-related public sentiment trends over a decade, identifying key trends in vaccine skepticism,

emotional intensity, and polarization. Our findings offer valuable insights for strategic public health communication, policy making, and intervention efforts.

3. **Financial analysis**: We demonstrate the applicability of LLM-enhanced stance detection methodologies in effectively classifying the managerial attitudes and assessing risks within corporate disclosures. Our approach and findings offer valuable tools for financial regulators, investors, and auditors by enabling the interpretation of complex financial texts.

4. **Prompt optimization**: We demonstrate that optimizing prompts by decomposing them into modular components, refining them iteratively with LLMs as optimizers, and adapting them to domain-specific linguistic and contextual features significantly improves LLM performance in stance detection. Our results establish prompt optimization as a critical methodological improvement for utilizing the full potential of LLMs in domain-adaptive stance detection tasks.

By integrating these methodological advancements, resource contributions, and empirical insights, this dissertation makes a significant contribution to stance detection research. It provides scalable and robust analytical frameworks that are crucial for navigating complex, nuanced, and polarized public discourses.

Some contributions from this work have been published in peer-reviewed conferences; others are under review, and we plan to submit additional manuscripts to conferences and journals as shown below:

1. Gyawali, N., Sirbu, I., Sosea, T., Khanal, S., Caragea, D., Rebedea, T., & Caragea, C. (2024, August). Gunstance: Stance detection for gun control and gun regulation. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 12027-12044). https://doi.org/10.18653/v1/2024.acl-long.650

2. Gyawali, N., Caragea, D., Caragea, C., & Mohammad, S. M. (n.d.). The shifting landscape of vaccine discourse: A massive dataset examining a decade of social media

posts on vaccines before and during the COVID-19 pandemic. *Manuscript submitted for publication to PLOS ONE, currently under review.*

3. Gyawali, N., Caragea, D., Vasenkov, A., & Caragea, C. (n.d.). Evaluating large language models for stance detection on financial targets from SEC filing reports and earnings call transcripts. *Manuscript submitted for publication to the Scientific Reports journal* (https://www.nature.com/srep/), *currently under review.*

4. Gyawali, N., Caragea, D., & Caragea, C. (n.d.). Modular prompt optimization for stance detection. *Manuscript in preparation; planned submission via ACL Rolling Review (ARR)*, Deadline: January 5th, 2026.

In addition to the publications included in this dissertation, the works below represent further contributions from my Ph.D. Although not part of the dissertation, they address cross-domain applications of machine learning and deep learning.

1. Gyawali, N., Hao, Y., Lin, G., Huang, J., Bika, R., Daza, L. C., ... & Liu, S. (2024). Using recurrent neural networks to detect supernumerary chromosomes in fungal strains causing blast diseases. *NAR Genomics and Bioinformatics*, *6*(3), lqae108. https://doi.org/10.1093/nargab/lqae108

2. Gyawali, N., Khanal, S., Caragea, D., Aziz, H. A., & Fitzsimmons, E. J. (2025). Predicting commercial motor vehicle crash severity in Kansas using explainable machine learning. *Journal of Transportation Safety & Security*, 1–32. https://doi.org/10.1080/19439962.2025.2540388

3. Taghian Dinani, S., Caragea, D., & Gyawali, N. (2023). Disaster Tweet classification using fine-tuned deep learning models versus zero and few-shot large language models. In: Gusikhin, O., Hammoudi, S., Cuzzocrea, A. (eds) Data Management Technologies and Applications. DATA 2023. Communications in Computer and Information Science, vol 2105. Springer, Cham. https://doi.org/10.1007/978-3-031-68919-2_4

4. Gyawali, N., Jha, A., Aziz, H. M. A., Caragea, D., & Fitzsimmons, E. J. (n.d.). Comparative geospatial analysis of weather-related impacts on crash frequency with explainable machine learning: A case for Kansas census tracts. *Manuscript in preparation; planned submission to a peer-reviewed conference or journal.*
   (*An initial version of this work was peer-reviewed and accepted for presentation at the Transportation Research Board (TRB) Annual Meeting 2024.*)

5. Jha, A. N., Hoehn, M., Mazin, A., Gyawali, N., Aziz, H. M. A., Caragea, D., & Fitzsimmons, E. J. (n.d.). Crash severity prediction using ExcelFormer, a transformer-based model for tabular data. *Manuscript in preparation; planned submission to a peer-reviewed conference or journal.*

6. Jha, A., Sabbagham, M., Gyawali, N., Aziz, H. M. A., Caragea, D., & Fitzsimmons, E. J. (n.d.). Graph neural network-based crash frequency prediction using road network and weather data: A case study from Kansas. *Manuscript in preparation; planned submission to a peer-reviewed conference or journal.*

Chapter 2

GunStance: Stance detection for gun control and gun regulation

2.1 Introduction

In the aftermath of the multitude of mass shooting events in the United States, topics regarding tighter gun regulations and control, expanding mental health services, and upgrading security in public places have been strongly debated over political campaigns on various social media platforms. People are generally on two opposing sides with respect to gun regulations and control: those who favor such regulations and those who strongly oppose them. Such differences in opinions on gun regulations have prevailed in society for over four decades, resulting in very little legislative success (Lin and Chung, 2020). Social media, especially X (formerly Twitter), has been a melting pot for such debates. It would be useful to automatically analyze the intricate dynamics, challenges, and complexities surrounding gun control topics.

Stance detection on social media is an emerging research area in opinion mining for various applications where sentiment analysis may be sub-optimal (ALDayel and Magdy, 2021). Stance detection is the task of automatically determining if an author of a text is *In-Favor*, *Against*, or *Neutral* towards a proposition or target (Mohammad et al., 2017).

While research on stance detection has seen significant progress in various domains, such as politics, healthcare, and finance (Conforti et al., 2020a; Ghosh et al., 2019; Glandt et al., 2021; Mohammad et al., 2017), there remains a notable gap in analyzing stance dynamics in the context of gun control and regulations on social media, despite the importance of this topic in contemporary society. This is mainly due to a lack of datasets labeled with respect to stance towards the controversial topics of gun control and regulation.

In this work, we aim to address this need by creating GunStance, a dataset of tweets collected during mass shooting events in the United States. Specifically, we focus on seven shooting events and identify two controversial, interconnected gun-related stance targets, a stronger stance target - "banning guns" and a softer stance target - "regulating guns." We collect tweets posted during those shooting events and filter tweets potentially relevant to each of the targets considered. Note that some tweets can express (either the same or different) stances towards both targets; therefore, the sets of tweets filtered for the two targets are not disjoint. Using Amazon Mechanical Turk, we annotate 2,700 tweets according to the author's stance toward the "banning guns" target, and 2,800 tweets according to the author's stance toward the "regulating guns" target. In addition, the GunStance dataset includes 7,334 unlabeled tweets for the "banning guns" target and 8,824 tweets for the "regulating guns" target. The unlabeled tweets can be utilized by semi-supervised learning approaches. More details about data collection and statistics are provided in Section 2.3.

In addition to assembling the GunStance dataset consisting of labeled and unlabeled tweets for the two targets, we first establish strong baselines for this dataset by employing supervised, semi-supervised, and zero-shot LLM-based approaches. To push the performance on our task and increase the real-world applicability of the stance detection models, we propose a hybrid SSL/LLM approach that utilizes unlabeled examples from a stream of data (i.e., posted as events unfold), while leveraging the zero-shot capabilities of LLMs to improve the model. Critically, compared to traditional SSL, our method incurs *minimal additional training costs*, due to our novel approach of identifying informative unlabeled examples, and *no inference costs*. Detailed information on our methods is provided in Section 2.4, while the results of the models are presented and discussed in Section 2.5.

To summarize, the main contributions of this work consist of providing the research community with a dataset, a novel low-cost hybrid SSL/LLM approach, and comparisons with meaningful baseline models for analyzing social media data with respect to stance towards controversial topics regarding gun regulations and control. Our dataset and proposed hybrid model are designed to help uncover prevalent arguments and opinions related to these topics and to foster a deeper understanding of the intricate dynamics surrounding gun control and regulations. By enabling the examination of diverse perspectives, our work has the potential to lead to advancements in addressing the challenges posed by gun control and regulation.

2.2 Related work

2.2.1 Stance detection datasets

The SemEval2016 dataset introduced by Mohammad et al. (2016, 2017) consists of tweets about US politics collected during the lead-up to the 2016 US Presidential election. Conforti et al. (2020a) presented a large dataset of approximately 50,000 tweets for stance detection. Villa-Cox et al. (2020) assembled a dataset for identifying stance in Twitter replies and quotes. Multi-target and multi-lingual datasets are contributed by several works, including (Glandt et al., 2021; Lai et al., 2020; Sobhani et al., 2017; Vamvas and Sennrich, 2020; Zotova et al., 2020). Several non-English datasets have also been published (Alturayeif et al., 2022; Küçük and Can, 2018; Lai et al., 2018; Lozhnikov et al., 2020). Furthermore, Allaway and McKeown (2020) and Xu et al. (2022) curated zero-shot and few-shot stance detection datasets. A summary of existing datasets is provided in Table A1 in the Appendix A.

While some previous works have incorporated the generic topic of "guns" as a target in their stance detection datasets (Cinelli et al., 2021; Sakketou et al., 2022), there has been limited differentiation between the intricate relationships surrounding the topics of "banning guns" and "regulating guns." Furthermore, prior studies have not targeted such topics, specifically in the aftermath of mass shooting events in the United States. As another

important difference, our dataset includes unlabeled tweets, which makes it possible to study semi-supervised and zero-shot/few-shot approaches, while most of the existing datasets only include labeled data. Therefore, our dataset provides a novel contribution in terms of stance targets included and approaches enabled.

2.2.2 Stance detection approaches

Support vector machines (SVM) with manually engineered features were established as strong baselines for the SemEval2016 Task 6 datasets (Mohammad et al., 2016, 2017). Subsequently, deep learning approaches, including recurrent neural networks (RNNs) and convolutional neural networks (CNNs), that incorporate both input and target-specific information using attention mechanisms (Augenstein et al., 2016; Du et al., 2017; Siddiqua et al., 2019; Sun et al., 2018; Zhou et al., 2017) outperformed the prior SVM strong baselines. With the advent of transformers (Vaswani et al., 2017) and bidirectional encoder representation from transformers (BERT) (Devlin et al., 2019) in NLP tasks, recent works (Li and Caragea, 2021; Slovikovskaya and Attardi, 2020) have investigated the use of BERT for stance detection. Ghosh et al. (2019) found BERT to be the best model overall for stance detection on the SemEval2016 Task 6.

In addition to utilizing target-specific information, several studies have demonstrated that incorporating auxiliary information, such as sentiment, emotion, Wikipedia knowledge, etc. can enhance the performance of stance detection models as compared to using only tweet/-target information (Fu et al., 2022; He et al., 2022; Hosseinia et al., 2020; Li and Caragea, 2021; Luo et al., 2022; Mohammad et al., 2017; Sun et al., 2019; Xu et al., 2020). Moreover, in scenarios where the labeled data for the target task is limited, various approaches have been employed to enhance performance. These include techniques such as weak supervision (Wei and Zou, 2019), knowledge distillation (Miao et al., 2020), knowledge graphs (Liu et al., 2021a), transfer learning using pre-trained models (Ebner et al., 2019; Hosseinia et al., 2020), distant supervision (Zarrella and Marsh, 2016), and domain adaptation from a source to a target task (Xu et al., 2020; Xue and Li, 2018).

Following prior works, to address the scarcity of labeled data, we propose a hybrid SSL/LLM approach that makes use of unlabeled data by leveraging the capabilities of LLMs. We compare the proposed hybrid approach with supervised, SSL, and LLM-based zero-shot strong baselines on the GunStance dataset.

Table 2.1: Examples of annotated tweets.

Tweet	**Target**	**Stance**
I am following reports of a shooting at #SixFlag in the #Chicago suburb of #Gurnee #Illinois. #GunControl NOW!	Regulating guns	*In-Favor*
Illinois has some of the strictest gun laws in the country, yet policymakers are scrambling to explain the series of missed warning signs and communication failures that led to the Fourth of July mass shooting in Highland Park.	Regulating guns	*Against*
One would wonder how this Highland Park shooting suspect got his guns with all the gun laws in IL to flag with his mental sickness?	Regulating guns	*Neutral*
The shooter in Tulsa bought an AR-15 just hours before the shooting; shooters in both Buffalo and Uvalde—both just turned 18—also bought their rifles shortly before killing. We need a moratorium on sales of AR/AK rifles for now; that will save lives.	Banning guns	*In-Favor*
With your logic, please explain why the Buffalo NY shooting happened if gun laws will stop bad guys shooting?	Banning guns	*Against*
Which of these laws would have stopped the Buffalo shooting or the Uvalde shooting? In the Uvalde shooting the door was open but was supposed to be closed. You can't legislate doors to work!	Banning guns	*Neutral*

2.3 GunStance dataset

Dataset collection and filtering: Our data collection process involved gathering tweets related to mass shooting incidents across different locations in the United States. Using Wikipedia pages that list mass shooting events in the United States in 2021[1] and 2022[2], respectively, we identified and focused on 7 mass shooting events (3 in 2021 and 4 in the

[1] https://en.wikipedia.org/wiki/List_of_mass_shootings_in_the_United_States_in_2021
[2] https://en.wikipedia.org/wiki/List_of_mass_shootings_in_the_United_States_in_2022

first half of 2022), each with at least 5 dead people. The events selected (shown in Table 2.2) cover a variety of locations/settings (including groceries, a parade, a school, massage parlors, etc).

Past tweets were retrieved in chronological order by using Twitter's v2 full-archive search endpoint[3]. For each event, we collected tweets posted the day of the event and up to eight weeks after the event. We started the search using a basic set of seed rules specific to the location of each event. As an example, the seed rule for the Robb Elementary School shooting in Uvalde, Texas was: "((uvalde OR texas) shooting) OR uvaldeshooting OR uvaldetexas". The seed rules for the other events are shown in Table A2 in Appendix A. The rule for an event was updated with the most frequent hashtags daily during the first week of the event and weekly during the following weeks. On each update of the rule, we identified the 15 most frequent hashtags during the update period (a day or a week), and added those hashtags as additional *OR* terms to the current rule. In addition, we also maintained a list of terms to ignore, specifically terms that were observed among the frequent hashtags but were too general and added noise to the collection (e.g., *mass, dead, breaking, update, usa, america*). The terms added with each update are shown in Table A3 in Appendix A. The tweets collected were capped at a maximum of 4,500 tweets per day and 25,000 tweets per week for each event. The final dataset collected included a total of 395,485 original tweets, replies, and quoted tweets (retweets were ignored and are thus not included in this count).

A preliminary analysis of the data showed that 77.48% of the tweets collected (i.e., 306,422 tweets) contained URLs (e.g., links to news articles). While the tweets containing URLs may express a stance towards the targets of interest in our study, given that they link to external resources (many of them related to media outlets), the stance expressed may not be the general public's stance but rather the stance of media outlets, political organizations, etc. As we aim to focus on detecting the general people's stance towards gun regulations, we removed such tweets.

We also removed tweets from a list of 2,834 Twitter handles corresponding to media outlets (e.g., PulpNews, NYDailyNews, ABC, etc.). The 2,834 handles were identified based

[3] https://developer.twitter.com/en/docs/twitter-api/tweets/search/introduction

on post frequency (as media outlets have a large number of posts, as shown in Figure A.2.1 in Appendix A) and/or the use of the "news" keyword in their username. Together, we filtered out 309,395 tweets out of the total 395,485 collected tweets, leaving us with a set of 86,090 tweets.

To identify tweets most relevant to the targets considered, we used *Sentence-BERT* (Reimers and Gurevych, 2019) to embed all tweets in the dataset, as well as the targets (expressed as "guns should be banned" versus "guns should be regulated") and ranked the tweets with respect to the targets based on cosine similarity. The similarity between these two targets themselves was 0.6980. We filtered out the least relevant tweets to these targets by applying a similarity threshold of 0.40. Consequently, we obtained 11,621 tweets for "banning guns" and 13,476 tweets for "regulating guns."

Dataset preprocessing: As part of the data preprocessing phase, we applied several steps to clean and enhance the dataset. These steps involved removing duplicate tweets, emoticons, and non-ASCII characters from the tweet texts. We also filtered out all user mentions (e.g., @username) from the tweets. We only kept English-language tweets, filtering out all non-English-language ones. For this, we selectively retained tweets with the language code "en" as specified in the tweet metadata during the crawling process. Figure A.3.2 in Appendix A shows the distribution of the number of words and length of tweets in our preprocessed dataset. A significant portion of the dataset consists of tweets containing 45-50 words, with the peak at 47 words. Similarly, the majority of the tweets have 270-280 characters, with the peak at 278 characters.

Data annotation: We annotated 2,700 randomly selected tweets for the "banning guns" target and 2,800 randomly selected tweets for the "regulating guns" target using the Amazon Mechanical Turk crowd-sourcing platform. We ran the annotation task in several iterations in order to develop our quality control steps. Initially, we annotated 100 examples internally (50 for "banning guns" and 50 for "regulating guns") to use as a qualification task. Then, we asked all annotators to annotate this set of 100 examples to measure how well they perform. Finally, we annotated all the data using three high-performing annotators and calculated the final label for a tweet using majority voting. We measured the inter-annotator agreement

using Krippendorff Alpha, and obtained an average value of α=0.63, indicating moderate agreement.

Table 2.1 shows examples of annotated tweets. We observe that some tweets have an implicit stance (e.g., "*In-Favor*" and "*Against*" examples for "regulating guns"), whereas others have an explicit stance (e.g., "*In-Favor*" and "*Against*" examples for "banning guns").

Table 2.2: Statistics on the labelled and unlabelled tweets for each shooting event and target. A = *Against*, F = *In-Favor*, N = *Neutral*, U = *Unlabelled*.

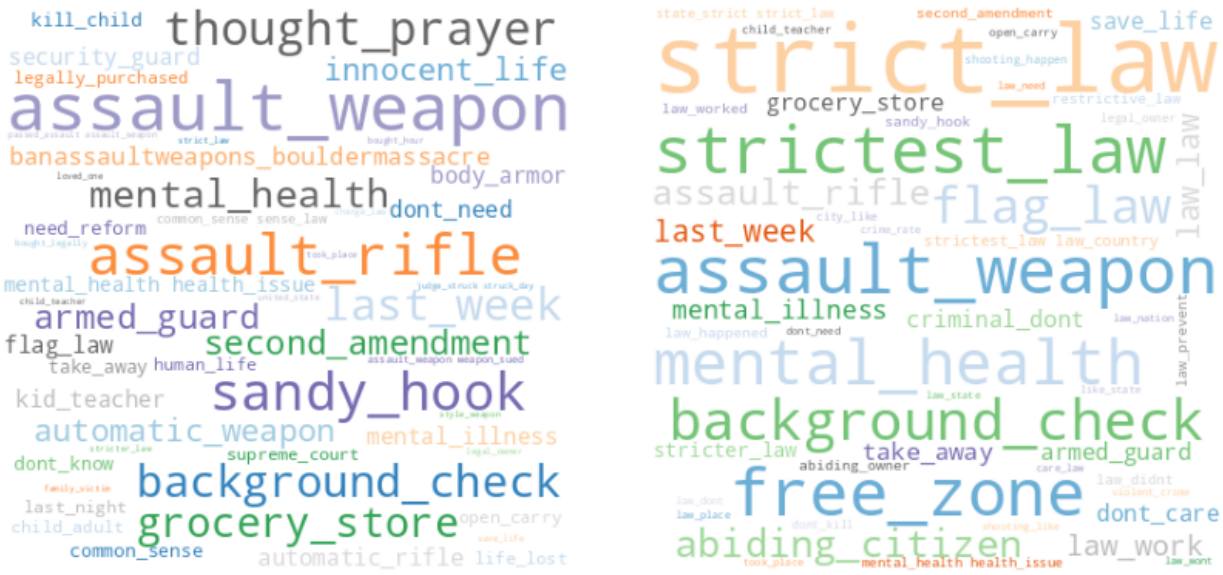
		"banning guns"				**"regulating guns"**			
Shooting event	**Date**	**A**	**F**	**N**	**U**	**A**	**F**	**N**	**U**
Atlanta, Georgia	03/16/2021	19	54	58	340	25	88	25	376
Boulder, Colorado	03/22/2021	73	117	126	572	86	188	78	704
San Jose, California	05/26/2021	20	24	21	221	13	35	9	224
Buffalo, New York	05/14/2022	330	354	477	3,215	373	518	342	4,033
Uvalde, Texas	05/24/2022	190	241	315	2,178	179	383	205	2,610
Tulsa, Oklahoma	06/01/2022	22	61	48	371	11	92	23	392
Highland Park, Illinois	07/04/2022	32	58	60	437	31	51	45	485
Total	—	686	909	1,105	7,334	718	1,355	727	8,824

Dataset statistics: Table 2.2 provides statistics about the labeled and unlabeled tweets for each target and each shooting event. For each event, the table shows the number of tweets annotated as *Against* (A), *In-Favor* (F), *Neutral* (N), as well as the number of *Unlabelled* (U) tweets. The total number of labeled tweets for "banning guns" is 2,700, while the number of unlabeled tweets is 7,334. The number of labeled tweets for "regulating guns" is 2,800, while the number of unlabeled tweets is 8,824.

Figure 2.1 shows bi-gram word clouds for "*In-Favor*" and "*Against*" stances with respect to the "banning guns" target. While many bi-grams are shared between the two word clouds, we find that people *In-Favor* of banning guns employ more terms such as "thought prayer," "don't need," and "common sense," indicative of their inclination towards advocating gun prohibition. Conversely, individuals *Against* banning guns employ more often terminology such as "strict law," "free zone," and "mental health." In particular, the term "strict law" is used to argue that despite strict gun laws in states such as Illinois and New York, instances of crime persist, thereby challenging the efficacy of banning guns as a viable solution to

prevent gun violence, as shown in the examples in Table 2.1.

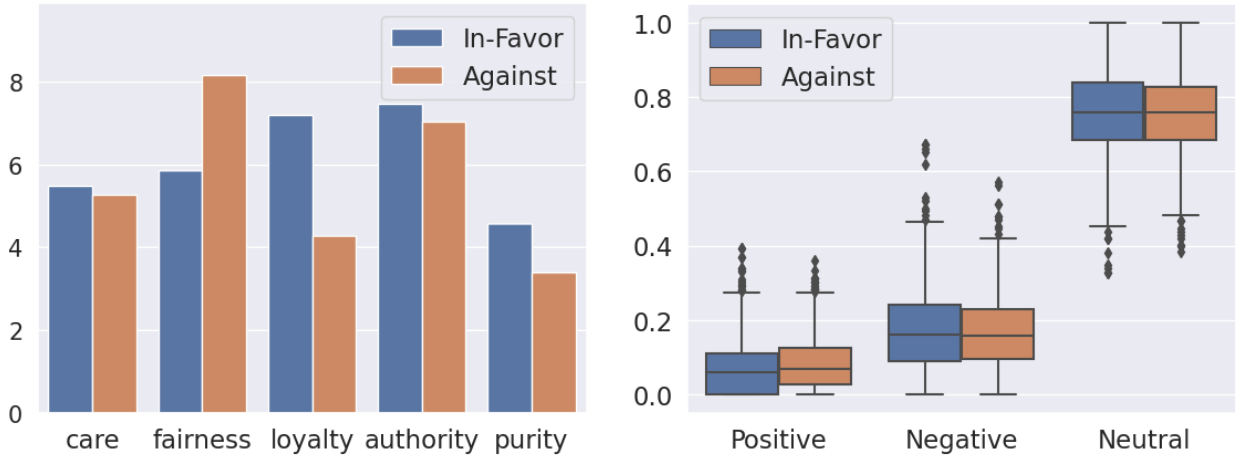
We also performed a lexical analysis of tweets pertaining to the "*Against*" and "*In-Favor*" stances of the "banning guns" target in terms of the five basic moral foundations (harm/-care, cheating/fairness, betrayal/loyalty, subversion/authority and degradation/purity) of the Moral Foundation Theory (Araque et al., 2020, 2022; Graham et al., 2013), as shown in the Figure 2.2 (Left). Each dimension is rated on a continuous scale ranging from 1 to 9 (1 for "harm" and 9 for "care"; 1 for "cheating" and 9 for "fairness"; 1 for "betrayal" and 9 for "loyalty"; 1 for "subversion" and 9 for "authority"; and 1 for "degradation" and 9 for "purity"). We find that individuals *In-Favor* of "banning guns" tend to express sentiments with a higher score than those *Against* "banning guns" in the dimensions of care, loyalty, authority, and purity. In contrast, those *Against* "banning guns" tend to articulate their viewpoints more in alignment with the dimension of fairness.



(A) *In-Favor* (B) *Against*

Figure 2.1: Bi-gram word clouds for (a) *In-Favor* and (b) *Against* stances on the "banning guns" target.

Employing VADER (Valence Aware Dictionary and sEntiment Reasoner) (Hutto and Gilbert, 2014), a lexicon-based sentiment analysis tool designed to capture sentiments in



(A) Five moral-foundation dimensions

(B) Sentiment scores

Figure 2.2: Scores for (A) the five basic moral-foundation dimensions and (B) overall sentiment in tweets expressing *Against* and *In-Favor* stances toward the "banning guns" target.

Table 2.3: Statistics of the labeled dataset splits in GunStance. "All" is the tweet count from all events; "Ban"=tweets about "banning guns"; "Reg."=tweets about "regulating guns."

		All		Boulder		Buffalo		Uvalde	
		Ban	Reg.	Ban	Reg.	Ban	Reg.	Ban	Reg.
Train	Against	228	243	25	32	110	128	64	56
	In-favor	304	449	39	62	118	168	81	132
	Neutral	368	244	42	24	159	115	105	69
Test	Against	228	247	24	29	110	124	63	64
	In-favor	303	433	39	59	118	171	80	118
	Neutral	369	252	42	29	159	116	105	73
Dev	Against	230	228	24	25	110	121	63	59
	In-favor	302	473	39	67	118	179	80	133
	Neutral	368	231	42	25	159	111	105	63
	Total	2,700	2,800	316	352	1,161	1,233	746	767

social media contexts, we performed a sentiment analysis on "*In-Favor*" and "*Against*" tweets pertaining to "banning guns", as shown in Figure 2.2 (Right). We did not find significant differences in sentiment between the "*In-Favor*" and "*Against*" stance categories. This result indicates substantial intricacies and nuances within this topic, thereby posing a significant challenge for machine learning models.

Benchmark subsets: To enable comparisons between our baselines and new models developed for the GunStance dataset, we randomly split each event in the dataset (using stratified sampling) into 33% training (`train`), 33% development (`dev`) and 33% test (`test`) subsets. We assemble benchmark subsets for the whole dataset by combining the `train`, `dev`, and `test` subsets, respectively, of the constituent events. Class distributions over each subset are shown in Table 2.3 under the "All" columns. The table also shows the class distributions for the three largest events in our dataset (specifically, "Boulder", "Buffalo" and "Uvalde"), as we also experiment with the datasets of these specific events in our study.

2.4 Methodology

2.4.1 Proposed approach: hybrid SSL/LLM

We leverage the Area Under the Margin Self Training (AUM-ST), an SSL model introduced by Sosea and Caragea (2022), and propose a novel approach by aiding AUM-ST (Sosea and Caragea, 2022) with ChatGPT during the training process. We call this model AUM-ST$_{+ChatGPT}$. As shooting events unfold, the web is constantly flooded with new streams of unlabeled data, which can be an important source of information to increase model performance and ensure up-to-date information is encoded into the models. Typically, SSL methods such as AUM-ST can use this stream of unlabeled data to improve model performance by adding the new data to the unlabeled set of examples. However, due to distribution shifts between the labeled and the recently collected unlabeled data, the model may struggle to generalize to the ongoing events. One alternative is to leverage the knowledge of LLM models, such as ChatGPT, by using them in a zero-shot manner. Unfortunately, using

Algorithm 1 AUM-ST$_{+LLM}$

Require: Labeled data $L = \{(x_1, y_1), (x_2, y_2), ...(x_n, y_n)\}$, unlabeled data $U = \{\hat{x}_1, \hat{x}_2, ...\hat{x}_m\}$; τ confidence threshold; and γ AvgMargin threshold.

1: Learn teacher model θ^t on labeled data minimizing the cross entropy loss

$$L_{\theta^t} = \frac{1}{n}\sum_{i=1}^{n} H(y_i, p(y|\pi(x_i); \theta^t))$$

2: Use the teacher model to generate hard pseudo labels for unlabeled examples

$$\hat{y}_i = \text{argmax}(p(y|\pi(\hat{x}_i); \theta^t)), \forall i = 1, \cdots, m$$

3: Train model θ^{AUM} on labeled and unlabeled examples, and monitor the training dynamics of unlabeled examples over $\mathcal{T}$ epochs using the Margin of each example averaged across the epochs: $AvgMargin(\hat{x}_i, \hat{y}_i) = \frac{1}{\mathcal{T}}\sum_{1}^{\mathcal{T}}[z_{\hat{y}_i} - max_{\hat{y}_i != j}(z_j)]$,

$z_{\hat{y}_i}$ and z_j are the logits corresponding to the pseudo-label $\hat{y}_i$ and the largest other logit.

4: Rank $\hat{x}_i$ based on $\phi(\hat{x}_i, \hat{y}_i) = Abs(AvgMargin(\hat{x}_i, \hat{y}_i))$ and select k unlabeled examples $U^{LLM} = \{\hat{x}_1, \hat{x}_2, ..., \hat{x}_k\}$ with low ϕ values.

5: Use an LLM model θ^{LLM} to generate pseudo-labels for examples in U^{LLM} (i.e. those examples that are perceived as ambiguous or hard-to-learn by the model) and overwrite the pseudo-labels generated by the teacher network with those generated by the LLM.

6: Train a student model θ^s on the labeled and high-confidence, high-average-Margin, and LLM-provided (i.e., the data pseudo-labeled using the LLM) pseudo-labeled data by minimizing the cross entropy loss:

$$L_{\theta^s} = \frac{1}{n}\sum_{i=1}^{n} H(y_i, p(y|\pi(x_i); \theta^s)) + \frac{1}{m} 1(max(p(y|\pi(\hat{x}_i) \geq \tau)) 1(AvgMargin(\pi(\hat{x}_i), \hat{y}_i) > \gamma) 1(\hat{x}_i \notin U^{LLM}) \sum_{i=1}^{m} H(\hat{y}_i, p(y|\Pi(\hat{x}_i); \theta^s))$$
$$+ \sum_{\hat{x}_i \in U^{LLM}} H(argmax(p|\pi(\hat{x}_i); \theta^{LLM})), p(y|\Pi(\hat{x}_i); \theta^s))$$

7: Use the student as a teacher and go back to Step 2

Table 2.4: Performance comparison for baseline models trained and evaluated independently for the "ban" and "regulate" targets. The models are trained using the combined training data of all events, and evaluated on the combined test data of all events. The best performance on each row is **highlighted**. For ECE, the lower the better.

	Zero-shot Models			Supervised Models							Semi-supervised Models			
	BART-NLI	**FLAN**	**ChatGPT**	**BERTweet**	**ATGRU**	**BiLSTM**	**GCAE**	**Kim-CNN**	**PGCNN**	**TAN**	**BERT-NS**	**FixMatch**	**AUM-ST**	**AUM-ST**$_{+ChatGPT}$
Target: Ban														
Accuracy	46.11	52.33	64.11	58.07	48.96	49.89	49.96	52.41	51.30	51.07	58.28	58.44	60.78	**66.38**
Precision	56.42	59.82	64.22	58.45	49.08	49.99	50.15	52.62	41.48	51.19	58.27	58.77	60.86	**66.32**
Recall	46.11	52.33	64.11	58.07	48.96	49.89	49.96	52.41	51.30	51.07	58.28	58.44	60.78	**66.72**
F1-Score	38.74	40.93	63.62	57.89	48.95	49.82	49.90	52.43	51.23	51.07	58.25	58.38	60.58	**66.50**
ECE	N/A	N/A	N/A	3.30	3.88	3.70	3.85	3.18	3.77	3.47	4.67	**3.08***	3.99	4.37
Target: Regulate														
Accuracy	42.81	56.65	65.02	63.71	53.51	58.33	55.22	59.33	57.44	57.15	64.10	64.52	65.77	**66.37**
Precision	57.27	52.92	65.34	63.40	53.11	58.11	54.66	58.57	56.61	56.54	64.50	64.17	65.40	**66.22**
Recall	42.81	56.65	65.02	63.71	53.51	58.33	55.22	59.33	57.44	57.15	65.09	64.52	65.77	**66.05**
F1-Score	38.61	47.28	64.67	63.09	53.08	58.13	54.57	57.99	56.73	56.64	64.27	63.90	65.45	**66.10**
ECE	N/A	N/A	N/A	4.08	3.67	3.94	4.29	3.17	**2.49***	3.18	3.73	3.80	3.42	3.75

LLMs at a large scale in these online, streaming setups is problematic as well, due to their high computational costs. We propose to leverage the high-quality predictions of LLMs in combination with AUM-ST to obtain an approach with low computational cost and better generalization performance. We show our approach in Algorithm 1 and highlight our changes to the vanilla AUM-ST in blue.

During SSL training, AUM-ST leverages AUM (Pleiss et al., 2020) to measure the fluctuation of predictions of the model on unlabeled data and to characterize the unlabeled examples based on the correctness of their pseudo-labels (Step 3). Examples with high AUM scores are likely to be pseudo-labeled correctly while low-AUM examples are likely incorrectly pseudo-labeled. On the other hand, examples with neither high nor low AUMs are examples whose pseudo-label correctness is uncertain. We argue that such examples are vital for model learning, and providing the model with correct pseudo-labels for these types of examples can significantly improve the performance. Since an AUM of zero indicates high uncertainty of pseudo-label correctness, we first propose to rank all unlabeled examples based on how close their AUM values are to an AUM of zero (Step 4). Then, we select uncertain examples and utilize ChatGPT as an external knowledge source to produce high-quality pseudo-labels for this particular category of unlabeled examples (Step 5). Concretely, at an arbitrary self-training iteration t, given the AUM of each unlabeled example (computed during AUM-ST training), we select a small percentage of unlabeled data with AUMs close to zero and use ChatGPT to obtain pseudo-labels for these examples. We then add these examples to the pseudo-labeled set for student model training (Step 6). We emphasize that the

examples selected for LLM pseudo-labeling change from one iteration to another, depending on the learning status of the model.

Note that our algorithm involves performing inference with ChatGPT on a very small fraction of the unlabeled set during the training process (only those that the teacher identifies as hard examples and has a high uncertainty on the label). Moreover, our method incurs no additional costs in practice since AUM-ST uses BERTweet as the base model.

2.4.2 Baseline models

We employ supervised, semi-supervised, and zero-shot models to establish baseline results for the proposed AUM-ST$_{+ChatGPT}$ model on our dataset. In the supervised learning setting, we only use labeled data to train the models. Similar to prior works in stance detection, e.g., (Ghosh et al., 2019; Glandt et al., 2021; Li and Caragea, 2021), we used **BiLSTM** (Schuster and Paliwal, 1997), **Kim-CNN** (Kim, 2014), **TAN** (Du et al., 2017), **PGCNN** (Huang and Carley, 2018), **ATGRU** (Zhou et al., 2017), **GCAE** (Xue and Li, 2018), and **BERTweet** (Loureiro et al., 2022) as our supervised baseline models.

Semi-supervised models are capable of leveraging unlabeled data to improve the performance of a teacher model. We utilized **BERT-NS** (Glandt et al., 2021; Hinton et al., 2015; Xie et al., 2020), **FixMatch** (Sohn et al., 2020), and **AUM-ST** (Sosea and Caragea, 2022) as SSL baselines.

For zero-shot (or few-shot) setups, a recent, powerful technique for addressing NLP tasks is to use large pre-trained language models (Brown et al., 2020). We employ **BART-NLI** (Lewis et al., 2020; Li et al., 2023b; Williams et al., 2018), **FLAN** (Wei et al., 2022a), and **ChatGPT** as zero-shot baselines.

A brief discussion of all the baseline models used is provided in Appendix A.4.

2.4.3 Experimental setup

The details of our experimental setup, including hyperparameters used for each of the baseline models, are presented in Appendix A.5. Each model was run three times using three

different random seeds. The results reported represent averages over the three runs.

We report the accuracy, precision, recall and F1-score as defined in Appendix A.5. In addition, we also report the expected calibration error (ECE). ECE (Naeini et al., 2015) measures how reliably a model's predicted probabilities reflect the true probabilities. The ECE values for a well-calibrated model should be low. Well-calibrated models increase the trustworthiness of the model's predictions, which is important to foster a deeper understanding of the intricate dynamics, challenges, and complexities surrounding gun control topics.

2.5 Results and discussion

We divide our experiments into three sets based on the data used to train and evaluate the models.

Experiment 1: In this experiment, we train and evaluate models independently for our two targets, "ban" and "regulate". Zero-shot models do not use any training data. Supervised models are trained using the combined labeled training data of all events. Semi-supervised models are trained using the combined labeled training data and the unlabeled data of all events. Subsequently, the trained models are evaluated on the combined test data of all events. Table 2.4 presents the Accuracy, Precision, Recall, F1-score, and Expected Calibration Error (ECE) metrics for the zero-shot, supervised, and semi-supervised baseline models considered, including our proposed SSL/LLM model, when trained/evaluated on "ban"/"regulate" train/test data, respectively.

AUM-ST$_{+ChatGPT}$, our semi-supervised model guided by ChatGPT, achieves the best performance across all metrics except for the ECE metric. Notably, the semi-supervised model FixMatch obtains the lowest ECE score for the "ban" target, while the supervised model PGCNN achieves the lowest ECE score for the "regulate" target. In terms of zero-shot models, ChatGPT outperforms both BART-NLI and FLAN significantly in all metrics.

For supervised models, we find a significant performance difference in terms of F1-Score between the non-transformer architecture-based models and the BERTweet model. This

Table 2.5: Performance comparison for the leave-one-event-out models trained for the "ban" and "regulate" targets and evaluated on "Boulder", "Buffalo" and "Uvalde", respectively. The best performance in each row is **highlighted**.

	Zero-shot Models			Supervised Models							Semi-supervised Models			
Event: Boulder														
Target: Ban														
	BART-NLI	**FLAN**	**ChatGPT**	**BERTweet**	**ATGRU**	**BiLSTM**	**GCAE**	**Kim-CNN**	**PGCNN**	**TAN**	**BERT-NS**	**FixMatch**	**AUM-ST**	**AUM-ST$_{+ChatGPT}$**
Accuracy	49.52	54.29	66.67	55.09	56.51	46.67	47.94	50.79	49.21	49.52	57.61	58.41	58.10	**67.40**
Precision	64.76	**72.10**	66.64	55.88	56.91	46.00	48.10	50.43	48.95	48.99	57.75	59.20	58.03	67.20
Recall	49.52	54.29	66.67	55.23	56.51	46.67	47.94	50.79	49.21	49.52	57.61	58.41	58.10	**67.03**
F1-Score	44.40	43.40	65.77	54.93	56.44	45.61	48.89	50.23	48.46	49.18	57.64	57.64	57.98	**67.10**
ECE	N/A	N/A	N/A	**4.45***	5.79	7.30	5.54	9.56	4.49	7.50	6.12	8.47	7.75	8.11
Target: Regulate														
Accuracy	53.85	58.12	64.96	67.52	54.70	60.68	58.11	62.39	57.26	61.53	67.24	68.09	**70.09**	68.54
Precision	**76.59**	46.55	64.68	66.87	55.38	58.89	56.03	60.59	55.41	60.24	65.78	67.58	69.23	68.99
Recall	53.85	58.12	64.96	67.52	54.70	60.68	58.11	62.39	57.26	61.53	67.24	68.09	**70.09**	68.65
F1-Score	50.38	48.52	63.87	66.95	54.56	59.44	55.75	59.74	55.46	60.65	65.38	66.05	**68.91**	68.78
ECE	N/A	N/A	N/A	7.38	6.92	6.68	7.36	**5.48***	11.32	8.69	7.36	9.27	6.10	7.44
Event: Buffalo														
Target: Ban														
	BART-NLI	**FLAN**	**ChatGPT**	**BERTweet**	**ATGRU**	**BiLSTM**	**GCAE**	**Kim-CNN**	**PGCNN**	**TAN**	**BERT-NS**	**FixMatch**	**AUM-ST**	**AUM-ST$_{+ChatGPT}$**
Accuracy	45.74	50.13	62.27	51.93	46.25	48.32	47.80	49.09	49.10	45.22	57.61	52.28	55.04	**63.10**
Precision	54.92	44.01	62.07	51.97	46.55	48.69	48.36	49.44	49.49	45.53	57.74	53.29	59.51	**64.74**
Recall	45.74	50.13	62.27	51.94	46.25	48.32	47.80	49.10	49.10	45.22	57.71	52.28	55.03	**63.12**
F1-Score	38.05	38.41	61.66	51773	45.93	47.70	46.23	48.68	48.18	44.90	57.62	51.53	53.65	**63.40**
ECE	N/A	N/A	N/A	5.50	6.89	6.81	5.83	**4.80***	5.49	6.01	6.63	6.65	7.71	8.99
Target: Regulate														
Accuracy	39.42	53.53	61.80	60.58	48.66	52.80	51.09	55.47	50.36	50.85	61.71	60.34	61.31	**63.54**
Precision	49.73	48.24	**64.49**	61.03	48.34	52.54	51.52	55.74	52.00	50.58	62.56	60.59	61.75	61.22
Recall	39.42	53.53	61.80	60.58	48.27	52.79	51.09	55.47	50.36	50.85	61.71	60.34	61.31	**67.50**
F1-Score	33.24	44.51	62.23	60.03	48.34	52.25	49.53	54.32	47.61	50.07	60.36	57.82	58.99	**64.50**
ECE	N/A	N/A	N/A	6.67	6.35	4.21	5.46	5.18	7.62	5.68	6.60	4.99	**3.57***	9.32
Event: Uvalde														
Target: Ban														
	BART-NLI	**FLAN**	**ChatGPT**	**BERTweet**	**ATGRU**	**BiLSTM**	**GCAE**	**Kim-CNN**	**PGCNN**	**TAN**	**BERT-NS**	**FixMatch**	**AUM-ST**	**AUM-ST$_{+ChatGPT}$**
Accuracy	46.37	53.63	66.94	54.83	43.15	43.95	42.74	49.20	47.98	47.18	56.98	53.76	58.06	**67.93**
Precision	51.53	**72.76**	67.63	54.82	44.26	46.04	41.23	50.79	48.31	48.62	57.05	54.20	57.80	68.20
Recall	46.37	53.63	66.94	54.83	43.15	43.95	42.74	49.19	47.98	47.18	56.98	53.76	58.06	**68.20**
F1-Score	38.30	42.70	66.87	53.79	42.04	43.85	39.98	48.49	47.24	45.65	56.99	52.50	57.20	**68.20**
ECE	N/A	N/A	N/A	5.19	6.10	11.32	8.46	6.41	9.30	5.03	6.57	5.67	**4.34***	6.31
Target: Regulate														
Accuracy	42.75	57.25	64.71	55.68	47.45	51.37	53.72	50.98	49.02	50.98	55.51	52.03	57.64	**65.32**
Precision	63.24	**71.15**	64.04	55.67	47.30	50.58	54.40	49.18	47.28	48.82	56.42	50.66	57.58	66.30
Recall	42.75	57.25	64.71	55.69	47.45	51.37	53.72	50.98	49.02	50.98	55.51	52.03	57.64	**65.18**
F1-Score	38.65	47.80	64.20	51.45	46.76	49.80	51.20	48.37	46.52	48.08	52.28	48.84	54.80	**65.80**
ECE	N/A	N/A	N/A	9.60	8.41	**5.72***	6.47	8.18	9.39	5.66	11.34	10.60	7.84	11.32

difference can be linked to the complexity inherent within our dataset, which presents a greater challenge for the supervised models to effectively model such complexities. In contrast, BERTweet was able to learn complexities present in our dataset better than the non-transformer supervised models. However, the zero-shot ChatGPT outperforms the best supervised BERTweet model, underscoring the power of LLMs, such as ChatGPT, in a zero-shot setting on a very challenging task.

Experiment 2: This experiment follows a leave-one-event-out setting. Models are trained for the "ban"/"regulate" targets independently using the training (labeled/unlabeled) data of all events except the held-out event, then tested on this event. The held-out events are "Boulder", "Buffalo", and "Uvalde". We did not consider other events as held-out because of their relatively small size.

Table 2.5 shows the results of the leave-one-event-out models for the "ban" and "regu-

late" targets evaluated on "Boulder", "Buffalo", and "Uvalde" events, respectively. Similar to Experiment 1, we find that the proposed AUM-ST$_{+ChatGPT}$ obtains impressive results across multiple events. Notably, AUM-ST$_{+ChatGPT}$ outperforms the vanilla AUM-ST by 9% F1 score on the "ban" target in the Boulder shooting and by almost 10% on the same target in the Buffalo shooting. Additionally, we emphasize that AUM-ST$_{+ChatGPT}$ consistently outperforms ChatGPT in all setups by approximately 2%, an impressive achievement, especially as the supervised baselines struggle on this challenging task. This result shows the value of the unlabeled data relevant to the task at hand, and suggests that our approach effectively leverages such unlabeled data (and the external knowledge captured through ChatGPT) to improve the generalization performance.

Experiment 3: In this experiment, we explore a cross-target setup, where we train the models on one stance target and test them on the other target. Specifically, we train the models on the training data of the "ban" target and evaluate their performance on the test data of the "regulate" target, and vice versa. This is a particularly difficult task, as differences between stances towards the "ban" and "regulate" targets can be very nuanced. To ensure good coverage with the unlabeled data, the semi-supervised models utilize the unlabeled data from both "ban" and "regulate" targets.

The results of this experiment are shown in Figure 2.3. As we can see from the figure, AUM-ST$_{+ChatGPT}$ outperforms both the supervised as well as other SSL methods for the "Ban→Regulate" setup and performs similarly with the other approaches for the "Regulate→Ban" setup.

2.6 Conclusions and future work

We presented GunStance, a stance detection dataset containing tweets pertaining to the controversial topics of "banning guns" and "regulating guns" in the aftermath of various shooting events in the United States. The dataset includes manually annotated tweets together with unlabeled tweets to facilitate supervised and semi-supervised learning. We

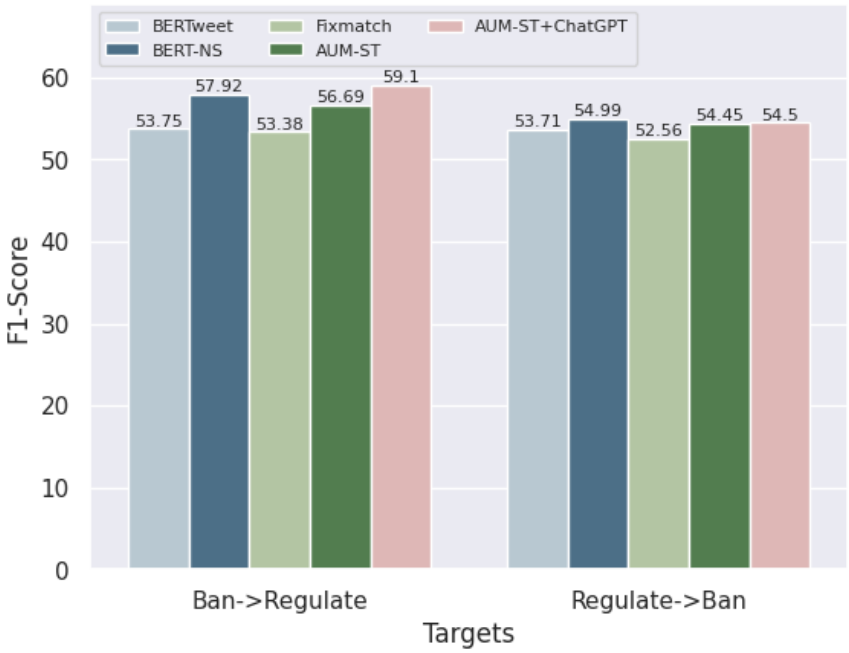


Figure 2.3: F1-scores of BERTweet, BERT-NS, FixMatch, AUM-ST, and AUM-ST$_{+\text{ChatGPT}}$ in the cross-target setup.

provide baseline results using established and newer supervised and SSL models for stance detection. We also utilize LLMs in a zero-shot setting and find that LLMs, such as ChatGPT, are able to understand the complexities in our dataset and outperform supervised baselines. Furthermore, we propose a curriculum-based semi-supervised approach that utilizes ChatGPT as an external knowledge source on a small subset of hard-to-learn unlabeled examples when the SSL model is less confident, pushing both performance and computational efficiency of SSL methods, such as AUM-ST.

The stance detection task addressed is not only timely (given the large number of shooting events), but also challenging. Our work has the potential to enable researchers to use machine learning models to analyze mass shooting data, gain insights into prevalent arguments and opinions, and develop a deeper understanding of the complexities surrounding gun control and regulations.

While there are other targets of interest in the context of mass shooting and gun control topics (e.g., buyback programs, background checks, mental health, etc.), we believe our work represents an important step towards understanding the general public's stance towards two

extremely important and interconnected gun control targets. Other related targets are left as part of future work.

An evaluation of the proposed AUM-ST$_{+ChatGPT}$ model on other datasets is beyond the scope of this current study, but it makes another excellent topic for future work. We also plan to focus on domain adaptation and transfer learning techniques that allow stance detection models to be adapted to different domains or targets. Finally, understanding similarities and differences between explicit and implicit stances in terms of LLM/SSL stance detection abilities is also of interest.

We make our dataset and code available on GitHub to spur further research in this area.

Chapter 3

The shifting landscape of vaccine discourse: A massive dataset examining a decade of English-language social media posts on vaccines before and during the COVID-19 pandemic

3.1 Introduction

Vaccines have been one of the most significant inventions in the history of mankind, and during the twentieth century, they helped eliminate most of the childhood diseases that were causing millions of deaths every year (Rappuoli et al., 2011). In the twenty-first century, vaccines still play a major part in safeguarding people's health from emerging infectious diseases, especially for vulnerable populations in low-income countries (Rappuoli et al., 2011).

On March 11, 2020, the World Health Organization (WHO) declared the novel coronavirus (COVID-19) outbreak a global pandemic, bringing the attention of the world to vaccines like never before. But with that, we have also seen a manifold increase in vaccine concerns and polarization of opinions. Thus, a deeper understanding of public perceptions of vaccines is crucial for devising effective communication strategies by health organizations.

X (formerly Twitter) has been one of the most engaging and influential micro-blogging platforms over the decades, allowing users to post and read short 280-character messages called posts (formerly tweets). As such, X has been regarded as a hub for social, political, and even health-related discussions. In this work, we study temporal patterns in English-language vaccine discourse on X, aiming to understand how such patterns have changed over the years, especially across the pre-COVID-19 (before 2020) and COVID-19 years (2020 to 2022). While extensive research exists on vaccine discourse, a critical gap remains in the availability of large-scale, high-quality datasets that track the evolution of public perceptions over time. Towards this goal, our study introduces a dataset consisting of 18.7 million meticulously curated English-language posts from a major social media platform, spanning a decade of discourse. In addition to introducing this large volume of unlabeled data, we also utilize theoretically grounded social cognition theory to examine the X users' perceptions regarding vaccines. This resource not only enables a comprehensive analysis of shifts in vaccine perceptions both before and during the COVID-19 pandemic but also serves as a valuable platform for advancing machine-learning methodologies. Specifically, the large volume of unlabeled data presents a promising opportunity to explore modern semi-supervised learning techniques, uncover deeper insights, and enhance automated annotation processes. Despite limitations on access to X posts since 2023, the vaccine-related X posts remain a significant and unique source of information to examine public perceptions. Other sources of information (such as surveys or even Reddit posts about vaccines) tend to be smaller, and while they are interesting in their own right, they do not make studying X posts redundant.

Our study draws on two complementary theoretical perspectives from social psychology and cognition theory to interpret large-scale English-language vaccine discourse on X.

Emotion dynamics: We apply emotion dynamics theory, which views emotions as dynamic, socially regulated processes that fluctuate across time and context. In psychology, emotion dynamics refers to the study of patterns of change and regularity in emotion (Hollenstein, 2015; Kuppens et al., 2010b). There is significant interest in understanding the dynamics of fluctuations in emotional state over time and how these changes differ among different people (Kuppens and Verduyn, 2017). While self-reported longitudinal data over a relatively short time period (Hamaker and Wichers, 2017) provides insights into emotion state and fluctuations, they only serve as a proxy for actual feelings. An alternative approach involves examining emotions through language usage. For instance, when experiencing happiness, people tend to use more happiness-associated words than usual, whereas during moments of anger, they are likely to use more anger-associated words (Rude et al., 2004). Although humans experience many emotions, several psychological studies highlight the importance of a select few (Ekman, 1999; Plutchik, 1991), e.g., the set of six Ekman emotions (*anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise*) (Ekman, 1999) or the set of eight Plutchik emotions that include Ekman's six emotions, as well as *trust* and *anticipation*. In this work, we focus on Plutchik's (Plutchik, 1991) eight emotions. For each emotion, we explore the extent to which that emotion exists across time: more trust or less trust. That is, when we explore the degree of trust, we explore the lack of trust (distrust/mistrust), the maximum amount of trust, and everything in between. This framework motivates our longitudinal analysis of shifting patterns in social media vaccine discourse.

Social cognition theory: We also draw ideas from social cognition theory and social psychology to understand the vaccine discourse on social media. Social cognition theory (Bandura, 2023) argues that people judge other people or social groups based on two key dimensions: *warmth* (which relates to friendliness, trustworthiness, sociability, fondness, reverence) and *competence* (which relates to ability, power, dominance, and assertiveness) (Abele et al., 2016; Bodenhausen et al., 2012; Fiske et al., 2002; Fiske, 2018; Koch et al., 2024). This framework is said to have developed because

of evolutionary pressures — early humans needed to quickly assess whether someone was a friend (warm/positive) or foe (cold/negative) and whether they were competent (powerful) or incompetent (weak). This theory is widely used to study inter-group perceptions, for example, how people from one country perceive individuals from other countries or how certain groups perceive homeless, immigrants, and so forth (Fiske, 2020). Essentially, these two dimensions – *warmth* and *competence* – can be used to create a 4-quadrant space to analyze how different target groups are perceived. Recently, the social cognition theory has been used widely to study not just perceptions of people and groups but also other subjects, such as brands (Fournier and Alvarez, 2012). In this work, we apply the theory to study the perception of English-speaking X users towards vaccines. Thus, we examine the social media discourse on vaccines along the two dimensions of social cognition theory, which can be summarized as: good–bad (*warmth*) and competent–incompetent (*dominance*). Such analysis provides insights into how language (stance, warmth/competence cues) towards the vaccines is framed in X and whether people believe they are effective or ineffective in achieving those outcomes (where one is located in this 2×2 competence–warmth space may determine what kind of public messaging is suitable for them). It should be acknowledged that perception towards individual vaccines can vary, with prominent vaccines sometimes shaping public opinions. In this work, we focus on the perception of English-speaking X users towards vaccines in general, although it would be interesting to explore sentiment towards individual vaccines. From our initial examination of the data, over 75% of the English-language vaccine posts on X during the COVID-19 years were related to COVID-19 vaccines.

We segment our analysis into two time periods: pre-COVID-19 vaccine discourse (2013–2019) and COVID-19 vaccine discourse (2020-2022) to examine how the pandemic has shifted vaccine-related discussions. To explore these shifts systematically, we focus on the following research questions (RQs):

RQ1. To what extent do English-language posts on X that mention vaccines use words as-

sociated with various emotions? How have the emotion word patterns changed before and during the COVID-19 pandemic years?

RQ2. From a social cognition perspective, how has the English-language discourse on vaccines on X changed over the years? Specifically, has there been an increase in the use of positive and warm words (suggesting growing trust and acceptance), or do we observe more competence-related words (highlighting perceptions of effectiveness and usefulness)?

RQ3. What approximate proportion of posts indicate a favorable stance towards vaccines, and what approximate proportion expresses opposition or skepticism? How have these proportions changed over time, especially in response to the COVID-19 pandemic?

RQ4. To what extent is vaccine opposition or skepticism characterized by untrustworthiness (low warmth) versus language that questions vaccine effectiveness (low competence)?

We focus on textual discourse rather than social media engagement metrics (likes, comments, and re-posts). While such engagement metrics may offer additional context, they do not directly capture how vaccines are framed within the social cognition dimensions. Our research aims to study vaccine discussions, rather than who discusses them or how content spreads. Engagement metrics can be ambiguous, as a *like* could signal agreement, sarcasm, or even disagreement in some cases. By focusing on linguistic content, we ensure our findings are directly linked to the framing and perception of vaccines, which is central to public discourse analysis. We study the emotions and stances at the population-level across thousands of English-language X posts, rather than classifying individual posts. For such aggregate analysis, lexicon-based methods are found to be empirically near-optimal while offering greater simplicity, interpretability, and substantially lower computational and carbon footprint compared to advanced models. Teodorescu & Mohammad (Teodorescu and Mohammad, 2023) show that when aggregating a few hundred instances per bin, lexicon-based methods produce very high correlation with the gold arcs (0.98 at bin=100), and the incremental gains from machine learning (e.g., Transformer models) at such aggregation levels

are minimal.

The four research questions systematically examine English-language vaccine discourse on X. RQ1 identifies and tracks emotion-related language. RQ2 builds on this by using social cognition theory to assess shifts towards more positive, competence-focused discourse. RQ3 focuses on users' stance, measuring support or criticism, while RQ4 explores negative discourse, distinguishing between emotional opposition and skepticism about effectiveness. This structured approach moves from general emotional trends to theoretical framing, explicit stance, and the nuances of skepticism, providing a comprehensive view of evolving vaccine perception.

To study these questions, we evaluate the emotional content of vaccine-related posts on X by examining the average frequency of emotional words used in posts from the pre-COVID-19 period (2013–2019) versus posts from the COVID-19 period (2020–2022). Using a Large Language Model (LLM) to assess stances towards vaccines, we analyze how discussions vary between pro-vaccine and anti-vaccine stances. We find that both positive and negative emotions increased during the pandemic, with notable rises in *surprise*, *trust*, and *anticipation*, reflecting heightened emotional responses and evolving perceptions in English-language X posts about vaccines. At the beginning of COVID-19, the discourse shifted towards more positive and competence-focused language but became more polarized with increased anti-vaccine sentiments towards the end of the pandemic. Anti-vaccine posts consistently used more negative words, highlighting persistent skepticism despite overall increased emotional engagement.

3.2 Related work

The analysis of social media posts, and especially posts from X, has been widely used for opinion mining for understanding public perception and opinions toward vaccines. Hu et al. (2021) looked into public opinion and perception of vaccines in the United States using spatiotemporal posts during the COVID-19-specific period. D'Andrea et al. (2019) used an automated system to infer trends in public opinion regarding the stance towards the

vaccination topic in Italian posts. A stance dataset towards vaccines during COVID-19 from posts in French, German, and Italian languages was collected by Giovanni et al. (2022). Similarly, various works carried out sentiment and opinion analysis on COVID-19 vaccines from social media posts, including countries like India and Italy (De Rosis et al., 2021; Praveen et al., 2021; Yousefinaghani et al., 2021). Works on emotion analysis from Stella et al. (2020) looked into emotion profiling of posts from Italian users during the COVID-19 pandemic lockdown for studying public feelings during unexpected events, whereas Alhuzali et al. (2022) studied the emotions and topics expressed on X during COVID-19 in the United Kingdom. More recent works curated data and studied sentiment, hesitancy, and dynamics towards vaccines during the COVID-19 (Alahmadi et al., 2025; Burwell et al., 2024; Lindelöf et al., 2023; Lyu et al., 2021; Mu et al., 2023; Qorib et al., 2023; Saleh et al., 2023; Sarirete, 2023). Poddar et al. (2024) studied the discourse of posts specific to anti-vaccines from 2018 to 2023. Osuji et al. (2022) conducted a U.S.-based survey to analyze the public perception of COVID-19 vaccine safety.

Table A1 provides additional details of works on sentiment and emotion analysis, along with topic modeling and stance detection on COVID-19 vaccine-related social media posts.

Lexical analysis has been applied in a wide variety of fields and adapted to languages other than English. For example, Aggarwal et al. (2020) used a valence-arousal-dominance framework to study the emotional expression differences between males and females from social media Reddit posts involving emotionally charged discourse during COVID-19. Hipson et al. (2021) used a computational linguistic approach to analyze posts crawled using terms such as "solitude," "lonely," and "alone" to understand how different people experience such emotions and the choice of language to discuss their experiences of these emotions.

Numerous studies have shown that the dimensions of warmth and competence have profound implications across a wide range of domains, including, interpersonal status (Swencionis et al., 2017), social class (Durante and Fiske, 2017), self-beliefs (Wojciszke et al., 2009), political and cultural perception (Fiske and Durante, 2016; Fiske et al., 2014), child development (Roussos and Dunham, 2016), and organizational processes such as hiring, employee evaluation, and resource (Cuddy et al., 2011). Research indicates that warmth significantly

influences the valence of interpersonal judgment, determining whether impressions are perceived as positive or negative.

Warmth is a core dimension in social perception and is often conceptualized as a subset of valence (Abele and Wojciszke, 2014; Oosterhof and Todorov, 2008). While valence captures a general emotional tone–positive or negative–warmth relates to perceived intent and encompasses two key facets: sociability (friendliness, likability) and trustworthiness (honesty, integrity) (Fiske et al., 2007; Leach et al., 2007). Together, these facets shape the judgment of a target's intentions, making warmth central to social and emotional evaluations.

While many studies focus exclusively on the COVID-19 period or specific emotions (and thus lack a comprehensive emotional analysis), our research examines over a decade of vaccine-related posts on X, leveraging human-annotated lexicons to assess sentiment and opinion shifts before and during the COVID-19 pandemic. Specifically, we analyze how emotional language associated with vaccines has evolved, investigating changes in the discourse from a social cognition perspective. We also study the proportion of posts indicating a favorable or skeptical stance towards vaccines and how these proportions have shifted over time, particularly during the COVID-19 pandemic. Finally, we assess the language that vaccine skeptics use, with a focus on low-competence language.

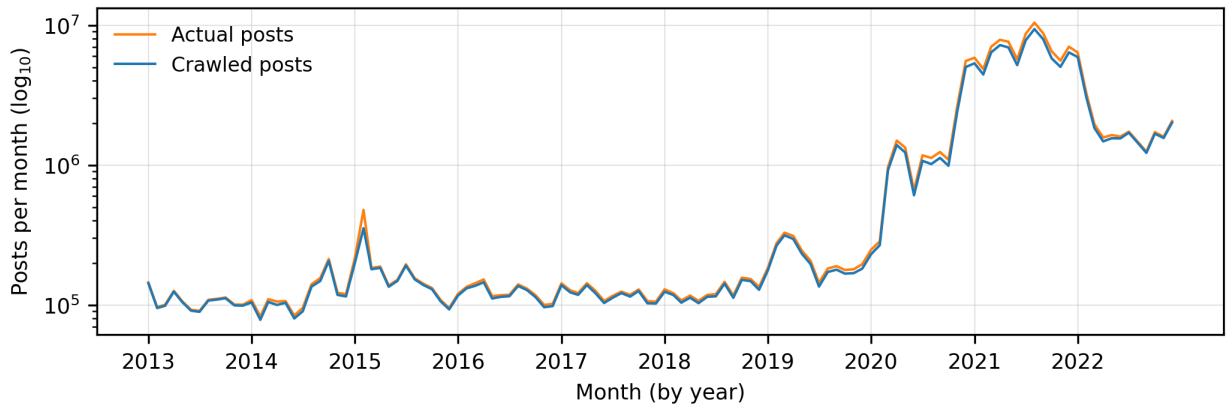
Data collection

Using X's full archive search API before they stopped academic API access on February 2023, we collected historical public posts for 10 years from the start of 2013 (2013/01/01) to the end of 2022 (2022/12/31), consisting of 7 years of posts before COVID-19 (2013–2019) and 3 years during the COVID-19 pandemic (2020–2022). We only collected English-language posts (excluding re-posts) using vaccine-related keywords. The specific keywords used are: *vaccine, vaccines, vaccinates, vaccinate, vaccinated, vaxxed, vaccination, vaccinations, #vaccine, #vaccines, #vaccinate, #vaccinated, #vaccination, #vaccinations, #vaccinated, #vaxxed, #vaccinate*. Fig 3.1A shows the volume of actual posts available to our search query (using X's counts API that allows retrieving the total posts count for a given query without actu-

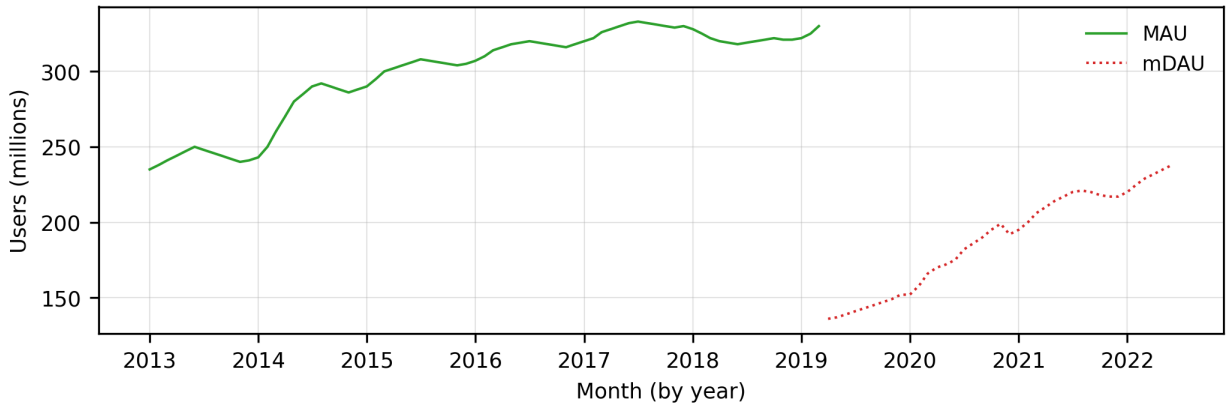
ally collecting the posts) and the volume of posts that we crawled, while Fig 3.1B shows the Monthly Active Users (MAU) and monetizable Daily Active Users (mDAU) gathered from official U.S. Securities and Exchange Commission (SEC) annual filings from Twitter before it went private. Twitter moved from reporting MAU to mDAU in April 2019. We find a significant rise in the total number of vaccine-related posts in the years following the start of the COVID-19 pandemic. Similarly, user engagement metrics show a gradual increase in MAU between 2013 and early 2019, followed by a steeper growth trajectory in mDAU from 2019 through the end of 2022.

After collecting the historical vaccine-related posts, we applied several preprocessing steps, as follows. We kept one post per user per day since the volume of posts we crawled was very large and also to reduce the influence of prolific user accounts on our analysis. We removed re-posts, emoticons, URLs, non-ASCII characters, and user mentions (e.g., @username) from the posts. Manual inspection of sample posts revealed that some of the crawled posts were related to a music band called "The Vaccines." To identify such posts, we used a Sentence-Transformer (Wang et al., 2020) fine-tuned on a large corpus of text to create embeddings for both the posts and the phrase "The Vaccines music band." We then filtered out posts based on their cosine similarity with the embedding of the phrase "The Vaccines music band," retaining only those with a similarity score smaller than 0.7. Approximately 2% of the total posts were filtered out. To assess the accuracy of this filtering procedure, we manually annotated a random sample of 200 posts (100 from the excluded set and 100 from the retained set). Treating references to the band as the positive class, 97% of the filtered out posts were correctly identified as referring to the music band (false-positive rate = 3%), while none of the retained posts referenced the band (false-negative rate = 0%), indicating a high level of classification precision.

We collected a total of 129M posts from 13M unique users. After preprocessing, we removed 85.59% of the total posts and 56.65% of the total unique users, leaving us with 18,730,502 posts and 5,872,259 unique users. This rigorous filtering ensures that our dataset is not only large in scale but also of high quality, making it a valuable resource for both



(A) Monthly volume of vaccine-related posts is shown in log_{10} scale. The orange line represents the total number of vaccine-related tweets posted ("actual posts"), and the blue line represents the tweets crawled by full-archive search ("crawled posts").



(B) Platform-level engagement metrics are plotted for the same period. The green line shows Twitter's Monthly Active Users (MAU), and the dotted red line represents monetizable Daily Active Users (mDAU), based on SEC filings prior to privatization. Twitter shifted from reporting MAU to mDAU in April 2019.

Figure 3.1: Comparison of Twitter activity related to vaccine posts and overall platform engagement over time. (A) Monthly volume of vaccine-related posts and (B) Platform-level engagement metrics during the same period.

current analysis and future research endeavors. Fig 3.2 shows the distribution of total words per post and total characters per post. The majority of posts contain 14-16 words, whereas the distribution of character counts per post is spread out, with the majority of posts being around 100-110 characters long. The dataset will be made available for research purposes to facilitate progress on tools that can help understand the vaccine discourse on social media.

As a control, we also curated tweets containing medical terms from 2015 to 2021 using the

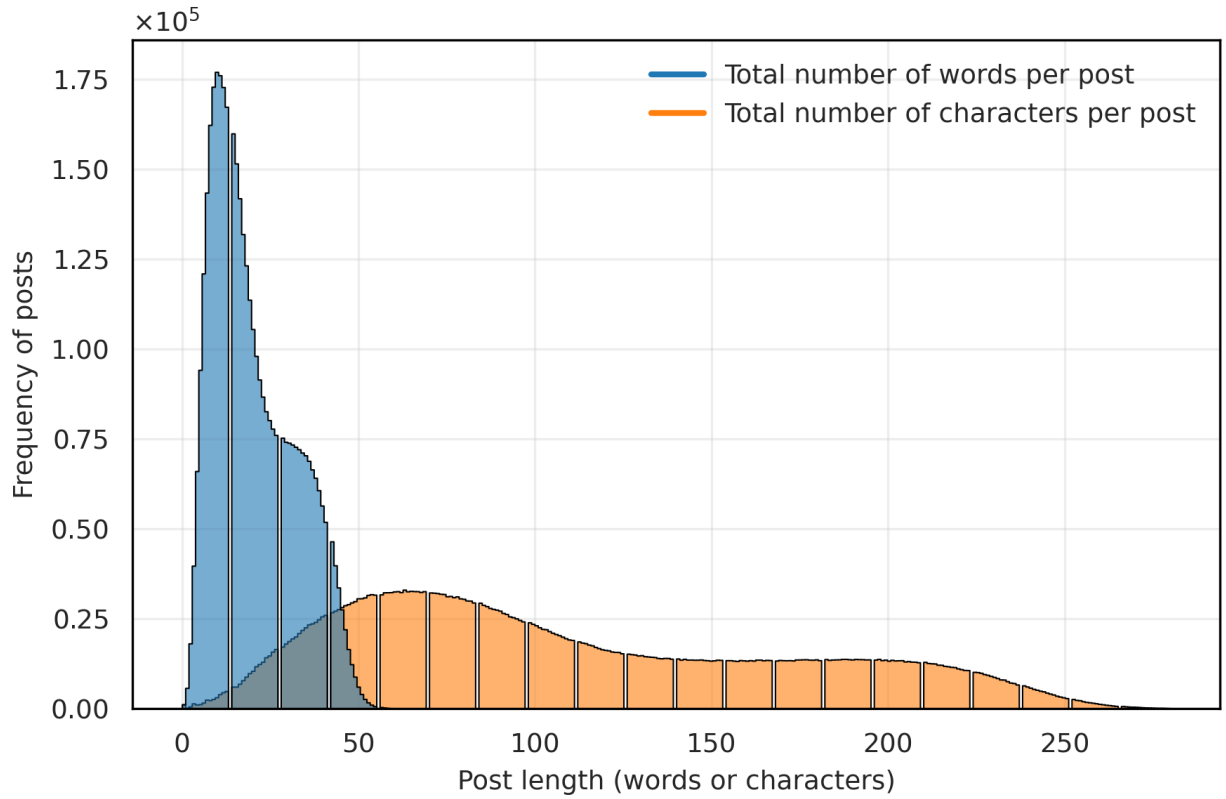


Figure 3.2: Distribution of word counts and post lengths. Distribution of the number of words per post and the total length of posts. The x-axis shows the count of words/length, and the y-axis shows the frequency of posts in units of 100,000.

TUSC-Dataset (Vishnubhotla and Mohammad, 2022). TUSC-Dataset contains more than 38 million tweets US and Canada from 2015-2021. From this dataset, we selected tweets that contain any of the keywords: *medicine*, *meds*, *medicines*, *antibiotic*, *antibiotics*, resulting in 100,000 tweets from 2015 to 2021.

3.3 Methodology

We employ the Utterance Emotion Dynamics (UED) metric framework (Hipson and Mohammad, 2021) that traces the emotional arc associated with utterances along the temporal axis. We rely on the NRC Emotion Lexicon (Mohammad and Turney, 2010, 2013), and the Words of Warmth Lexicon (Mohammad, 2018, 2025) to determine the UED metrics. To detect stance towards vaccines, we employ large language models (LLMs), specifically a Llama

model. The UED metrics, emotion lexicons, as well as the process used for LLM-based stance detection, are described in what follows.

3.3.1 Utterance Emotion Dynamics (UED) metrics

Home base: The home base is defined as a subspace of high-probability emotional state space where a speaker is most likely to be found (Kuppens et al., 2010a) and can be utilized to analyze utterances. For instance, at any given point in an individual's narrative, their location in the warmth–competence space may be determined by computing the averages of the warmth and competence scores of the words within a small window of utterance. The trajectory followed by this position over time represents the individual's emotional arc or trajectory. The home base is defined as the subspace where the individual is most likely to be located.

In the one-dimensional case, for example, warmth (w) or competence (c), the home base refers to the subspace associated with the most common average warmth or competence scores, respectively (Kuppens et al., 2010a). Mathematically, this band is defined as the lower and upper bounds of a confidence interval, as shown below for warmth:

$$\overline{w} \pm t_{(1-\alpha,N-1)}\sqrt{\frac{\sigma^2}{N}} \tag{3.1}$$

where $\overline{w}$ is the mean of w, t is the t-distribution, N is the number of components in w, α is the desired confidence (e.g., 68% –one standard deviation away from the mean), and σ^2 is the variance of the warmth values in w.

In the two-dimensional warmth–competence space, the home base is bounded within an ellipse defined by:

$$\frac{w_i - \overline{w}}{\psi\lambda_1} + \frac{d_i - \overline{d}}{\psi\lambda_2} = 1 \tag{3.2}$$

where $\overline{w}$ and $\overline{c}$ represents the means for warmth and competence, respectively, ψ denotes the critical χ^2 value for the desired confidence range, while λ_1 and λ_2 represent the eigenvalues of the covariance matrix. The values in the denominator of the two terms correspond to

the major and minor axes, representing the two diameters of the ellipse. The coordinates w and c that satisfy the equality in Eq (3.2) are the boundaries of the ellipse, whereas any set of coordinates for which the expression on the left of Eq (3.2) is smaller than 1 is located within the ellipse. Consequently, we can compare the home base ellipses among different individuals (or groups) in terms of their location in the warmth–competence space and their size based on the major and minor axes of the ellipses.

Emotional Variability (EV): Emotional variability measures how a speaker's emotional state changes over time. According to Krone et al. (2018), in a one-dimensional case, EV is defined as the Standard Deviation (SD) of the dimension considered, as shown below for warmth:

$$SD(w) = \frac{\sum_{i=1}^{N}(w_i - \overline{w})^2}{N} \tag{3.3}$$

In the two-dimensional case, EV is defined as the average of the standard deviations of w and c. The above metrics collectively capture the temporal emotional characteristics of an individual's utterances. A more detailed discussion of these metrics and their use for lexical analysis has been provided by Hipson and Mohammad (2021). We performed a similar lexical analysis utilizing the UED framework with warmth and dominance (competence) dimensions to study emotional expressions and opinions about vaccines, both in general over the 10 years and in the periods before and during the COVID-19 pandemic, using posts from the years 2013 to 2022.

3.3.2 Emotion lexicons

The UED metrics described above are determined using two existing word-emotion association lexicons: (1) the NRC Emotion Lexicon (Mohammad and Turney, 2010, 2013) and the Words of Warmth Lexicon (Mohammad, 2025).

The NRC Emotion Lexicon comprises about 14,000 commonly used English words and their associations with eight emotions (*anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, and *trust*), along with two sentiments (*positive* and *negative*) that are annotated manually using crowd-sourcing. The Words of Warmth Lexicon includes a list of more than

26,000 English words, each assigned a normalized score between 0 and 1 along the arousal, dominance (competence), trust, sociability, and warmth dimensions.

As we used vaccine-related words to crawl the posts in our dataset to study the shifting landscape of English-language vaccine discourse on X, we removed the entries in the lexicon that are morphological variants of the word *vaccine* (specifically, *vaccine* and *vaccination*). We also removed entries related to physical and mental illness (e.g., flu, polio, amnesia, etc.). This is to avoid the potential bias caused by the high frequency of these variants in our dataset.

It is important to note that the lexicons provide probable emotion associations without considering the contextual influence of the neighboring words within the target text. However, given that the majority of words usually possess a dominant primary sense (Kilgarriff, 2006), and the metrics capture emotion associations from a large number of words, this approach proves to be effective.

3.3.3 Stance detection

LLMs have been effective for stance detection in social media posts due to the vast amount of data they are trained on (Gambini et al., 2024). To identify the stance of posts about vaccines, we use the Meta's Llama 3.3 instruction-tuned model with 70 billion parameters (Dubey et al., 2024) and provide a prompt that asks the model to classify each post with respect to inferred stance of the poster towards vaccines as "favor," "against," or "neither of the two inferences can be reasonably made" (see Appendix B.1 for the exact prompt). Due to the large volume of posts in our dataset, we randomly sampled 2,000 posts per month across all years (a total of 240,000 posts) to be annotated by the LLM, as annotating the entire dataset would be very expensive both in terms of time as well as computational resources.

Our study goes beyond basic quantitative analysis by applying a theoretically grounded social cognition theory to examine how vaccines are perceived along the dimensions of warmth (positive-negative emotions) and competence (perceived effectiveness). In addition, we utilize

LLMs for longitudinal tracking of vaccine-related shifts from 2013–2022, identifying key trends before and during the COVID-19 pandemic. Using LLMs, we conduct scalable stance detection to classify attitudes toward vaccines and analyze emotional tone and perceived efficacy to understand changes in vaccine trust. By integrating social cognition theory, LLM-based stance detection, and longitudinal analysis, our study offers a deeper understanding of the evolving English-language vaccine discourse on X and could potentially have important implications for public health communication in English.

3.4 Results and discussion

In this section, we discuss the results and findings of our research questions.

RQ1. To what extent do English-language posts on X that mention vaccines use words associated with various emotions? How have the emotion word patterns changed before and during the COVID-19 pandemic years?

Figure 3.3 shows the emotion-word density of positive and negative sentiment words from the NRC Emotion Lexicon per month across various years. Darker lines show the three-month rolling average of emotion-word density, and lighter lines represent the observed monthly values. Emotion-word density is calculated as the total number of emotion-related words used in a month divided by the total number of words posted in that month. We find that, during the pre-COVID-19 years, the negative word usage varied more significantly as compared to the positive word usage, and English-speaking X users generally used more negative words as compared to positive words. Unlike any of the pre-COVID-19 years, the usage of negative words dropped significantly during early 2021, even below the usage of positive words, when vaccines were being developed and released for the public (see Appendix B.2 for the COVID-19 vaccine development timeline). However, we see a significant divergence from late 2021 onwards, where the usage of negative words increased rapidly, while positive word usage saw a declining trend. This may suggest vaccine hesitancy and discussions about the side effects of COVID-19 vaccines on X social media. The trend towards more negative language highlights challenges faced by public health officials and scientists in maintaining

public trust in vaccines.

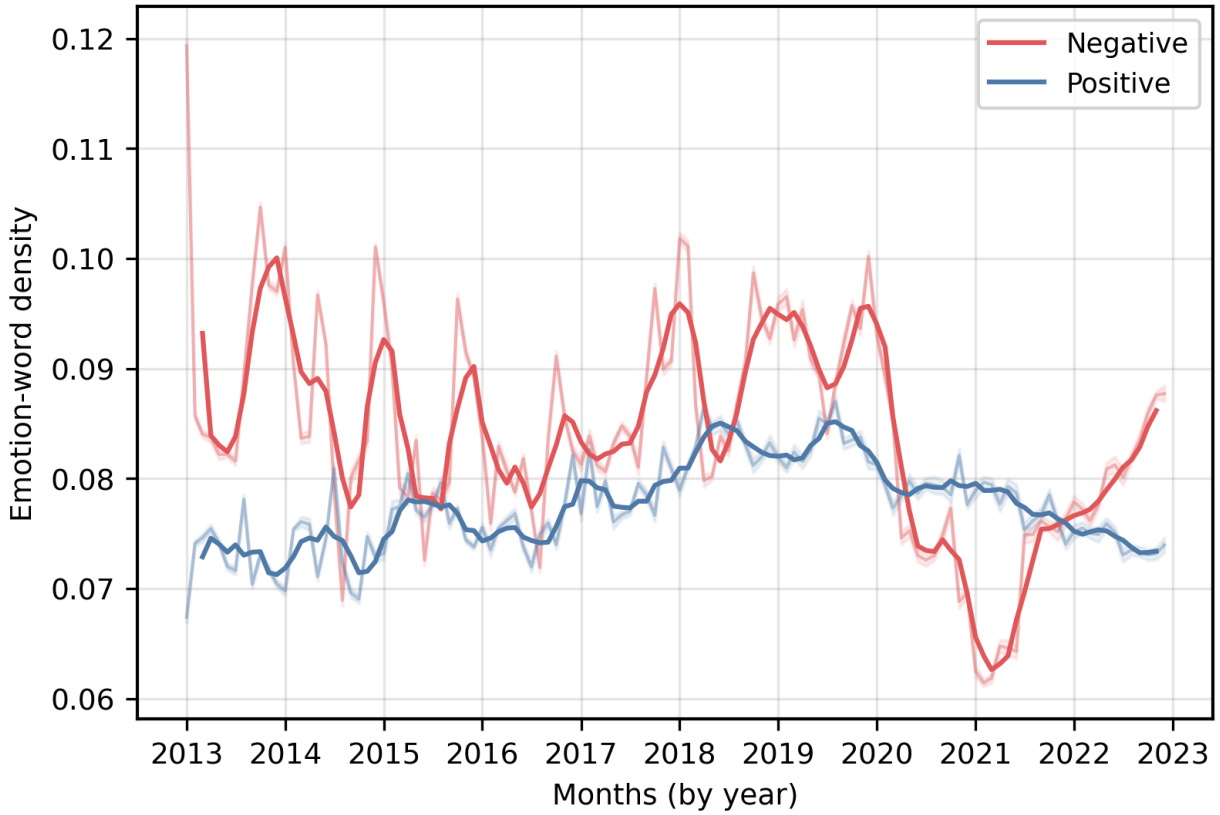


Figure 3.3: Monthly trend in emotion-word density of positive and negative sentiment words from 2013 to 2022. Darker lines show the three-month rolling averages of emotion-word density, and lighter lines represent the observed monthly values. Emotion-word density is calculated as the total number of emotion-related words used in a month divided by the total number of words posted in that month.

Figure 3.4 shows the monthly trends in emotion-word density during the COVID-19 years (2020-2022). The emotion-word density is calculated as the total number of emotion-related words used in a month divided by the total number of words posted in that month. Fear starts high in early 2020 and drops significantly at the beginning of 2021 before gradually increasing again in late 2021 and throughout 2022. This mirrors the trajectory of the pandemic: initial fear of the unknown, followed by more information and control measures, including the approval of COVID-19 vaccines, then renewed concerns with variants and long-term impacts of vaccines. Anticipation was higher at the beginning of 2020 but gradually decreased over 2021 and 2022, suggesting people were highly anticipating the vaccines, but

the anticipation decreased as vaccines became available after 2021. Other emotions like anger, disgust, joy, surprise, and trust remained relatively stable, suggesting these were not the dominant emotional responses to the pandemic.

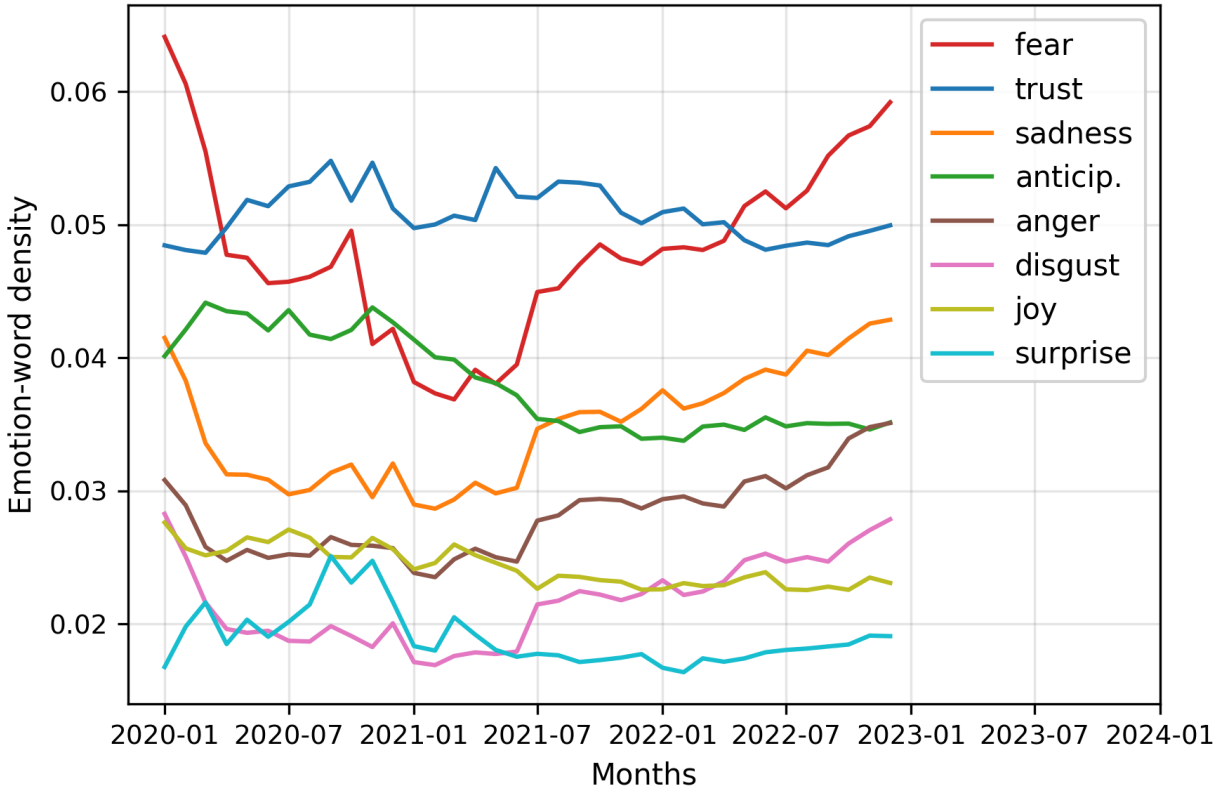


Figure 3.4: Monthly trends in emotion-word density during the COVID-19 pandemic (2020–2022). Emotion-word density is calculated as the total number of emotion-related words used in a month divided by the total number of words posted in that month.

Table 3.1 compares the emotion-word density of various emotions across all pre-COVID-19 years (2013–2019) and COVID-19 years (2020–2022). As before, emotion word density is calculated as the total number of emotion-related words used in a month divided by the total number of words posted in that month. The mean and standard deviation for each emotion are calculated using monthly word densities in pre-COVID-19 years and COVID-19 years. Negative emotion word usage decreased during the COVID-19 period, by 13.22% (from 0.0876 to 0.0760). However, we didn't find a statistically significant change in positive emotion word usage. We saw significant increase ($p < 0.001$) in surprise (from 0.0156 to 0.0190, i.e. 21.66%), and trust (from 0.0459 to 0.0508, i.e. 10.64%). However, other

emotions showed a decrease in COVID-19 years, notably, disgust (which decreased from 0.0301 to 0.0217, i.e., 27.81%) and fear (which decreased from 0.0632 to 0.0481, i.e., 23.89%), and also sadness (which decreased from 0.0406 to 0.0348, i.e., 14.23%), joy (which decreased from 0.0276 to 0.0243, i.e., 11.87%), and anger (which decreased from 0.0292 to 0.0281, i.e., 3.85%). Although the mean value of anticipation emotion increased, the result was not statistically significant.

Table 3.1: Emotion-word density of various emotions during pre-COVID-19 (2013–2019) and COVID-19 (2020–2022) periods, with percent change (% Change), and p-values (Mann-Whitney U test). Emotion-word density is calculated as the total number of emotion-related words used in a month divided by the total number of words posted in that month.

Emotion	**Pre-COVID-19**		**COVID-19**		**% Change**	**p-value**
	Mean	**SD**	**Mean**	**SD**		
Negative	0.0876	0.0085	0.0760	0.0079	-13.22	<0.001
Positive	0.0777	0.0047	0.0770	0.0025	-0.84	0.5344
Anger	0.0292	0.0023	0.0281	0.0031	-3.85	<0.05
Anticipation	0.0371	0.0020	0.0381	0.0037	2.65	0.5766
Disgust	0.0301	0.0045	0.0217	0.0032	-27.81	<0.001
Fear	0.0632	0.0076	0.0481	0.0067	-23.89	<0.001
Joy	0.0276	0.0019	0.0243	0.0015	-11.87	<0.001
Sadness	0.0406	0.0047	0.0348	0.0044	-14.23	<0.001
Surprise	0.0156	0.0018	0.0190	0.0021	+21.66	<0.001
Trust	0.0459	0.0039	0.0508	0.0019	+10.64	<0.001

Note: Percent change is based on the difference in mean values between COVID-19 and pre-COVID-19 periods, calculated as ((COVID-19 mean – Pre-COVID-19 mean) / Pre-COVID-19 mean) × 100. Values with $p < 0.05$ and $p < 0.001$ are considered statistically significant.

To conclude the analysis related to RQ1, the overall decrease in negative emotions suggests that vaccine discussions became slightly less negative during the pandemic. Higher usage of trust words suggests growing confidence in vaccines and expectations surrounding their development and distribution. The rise in surprise emotions could be related to rapid developments in vaccine research, policy changes, or unexpected events during the pan-

demic. The decrease in joy words could signal a dampened expression of happiness during the crisis. Finally, minimal change in positive and anticipation emotions could imply that expression of positivity and forward-looking sentiment remained relatively stable despite the upheaval. Overall, these trends suggest that while general negativity decreased, emotional expression was diversified–marked by greater trust and surprise–capturing both adaptation and uncertainty towards vaccines in the COVID-19 years.

RQ2. From a social cognition perspective, how has the English-language vaccine discourse on X changed over the years?

From a social cognition perspective, the discourse on vaccines has shifted over time, specifically during the COVID-19 pandemic. Figure 3.5 compares "home bases" for pre-COVID-19 and COVID-19 periods. We observe that these ellipses have different shapes: the pre-COVID-19 ellipse (red) is less widespread in the competence space, as compared to the COVID-19 ellipse (blue). This suggests that during pre-COVID-19, there was not much polarization around competence: that is, discussions around the effectiveness/competence of the vaccines were relatively stable. However, there was marked variability around competence during the COVID-19 years.

Figure 3.6 shows the comparison of warmth and competence dimensions. Figure 3.6A shows changes in warmth and its two distinct components - trust and sociability - over time from 2013 to 2022. We find a stable trend in warmth during pre-COVID-19 periods, whereas the usage of warm words increased in the COVID-19 period from 2020 to early 2021 before decreasing consistently until the end of 2022. We see a similar trend in the trust and sociability aspect of the warmth dimension. Interestingly, the usage of the warmth words remained consistent in control tweets containing medical terms, even during the COVID-19 period. This suggests that the public trust in vaccines may have increased during the early stages of the pandemic, but declined once vaccines became readily available. Several factors could have contributed to this shift, including concerns about the rapid development and roll-out of the vaccines. Figure 3.6B shows the change in competence over time. We find that competence increased during the first year of COVID-19 before decreasing consistently after early 2020. In the control medical tweets, we find that competence remains consistent during

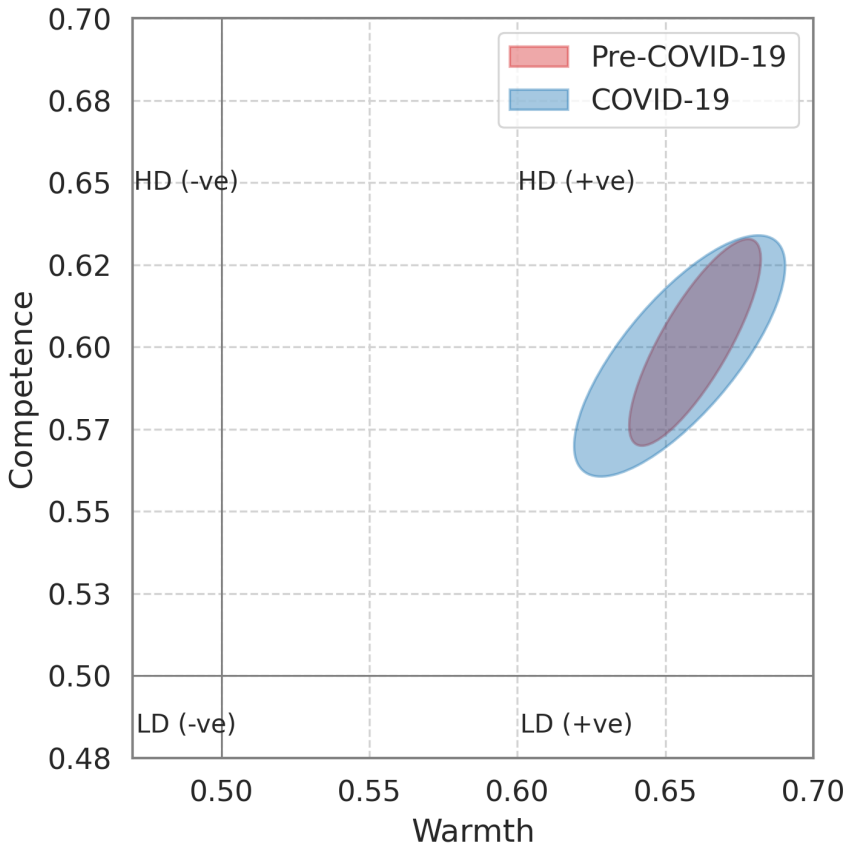
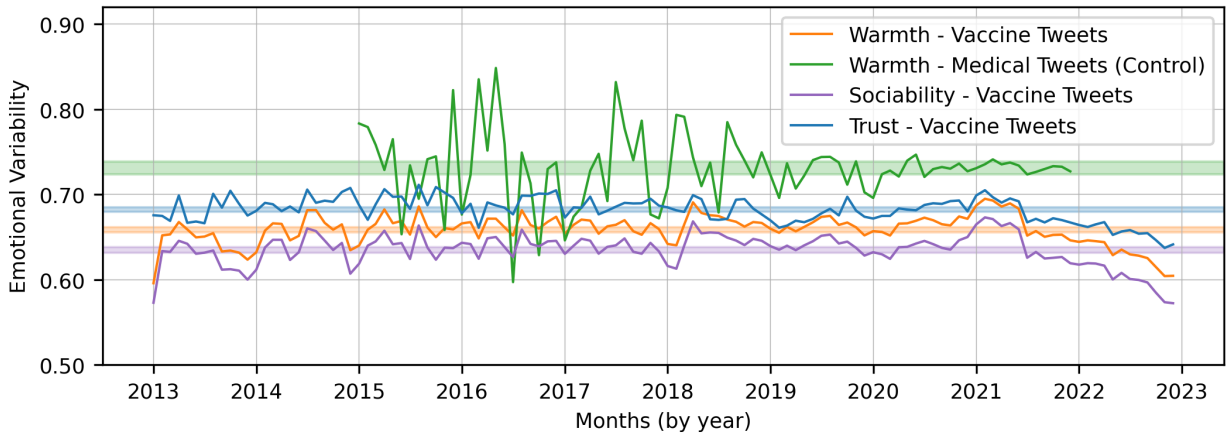


Figure 3.5: Comparison of vaccine-discourse "home bases" before and during COVID-19. The red ellipse represents home bases for pre-COVID-19 years and the blue during the COVID-19 years in the warmth–competence space.

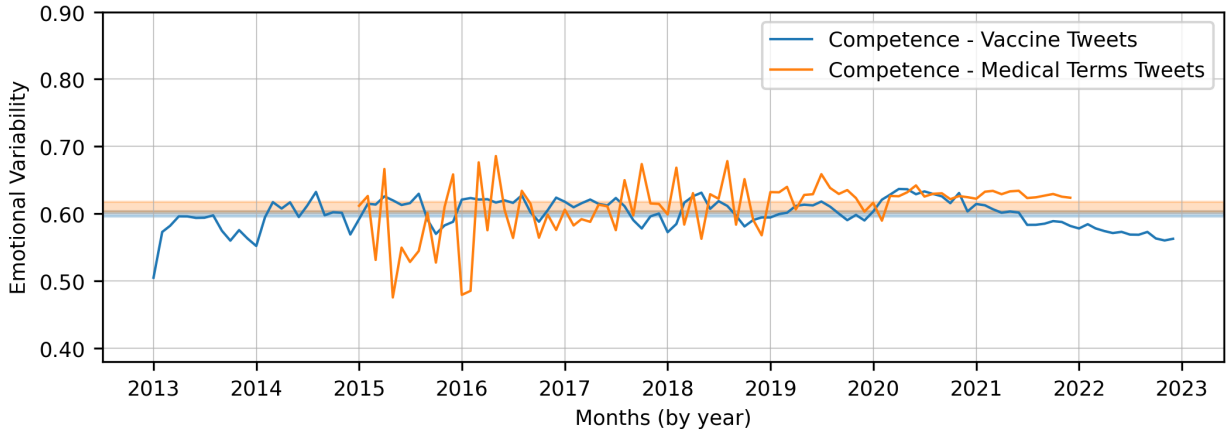
the COVID-19 years. This decrease in competence suggests heightened uncertainty and a lack of control. As the pandemic progressed, people shifted towards using more passive or cautious language. Additionally, public discourse often emphasized vulnerability, empathy, and shared hardship, which could have further contributed to the use of less competence-related words.

RQ3. What approximate proportion of posts indicate a favorable stance towards vaccines, and what approximate proportion expresses opposition or skepticism? How have these proportions changed over time, especially in response to the COVID-19 pandemic?

As mentioned above, we used the Llama 3.3 model to answer this question. To assess the model's accuracy, we manually annotated 500 posts and used the model to predict those posts



(A) Warmth and its two components—trust and sociability—over time.



(B) Competence over time.

Figure 3.6: Trends in warmth (trust, sociability) and competence language in vaccine discourse from 2013–2022. (A) Warmth and its two components—trust and sociability—over time. (B) Competence over time. Medical-term tweets serve as a control. The shaded bands represent the home bases for each dimension.

5 times. The average agreement between Llama 3.3 annotations and human annotations over the 5 runs was 66.32%, with a standard deviation of ± 1.94%. Note that random guessing will result in an accuracy of 33.33%. Table 3.2 shows the per-class precision, recall, and F1-score for each stance category, along with overall accuracy and averaged metrics for classification using the Llama 3.3 model. The model demonstrated strong performance in identifying both "favor" and "against" posts, with high recall, indicating a reliable detection of clear positive and negative stances. The temperature parameter in large language models (LLMs) modulates output randomness by shaping the sampling distribution. Lower temperatures

produce more deterministic text, whereas higher values enhance variability and creativity by increasing the likelihood of less probable tokens (Renze, 2024). We tested temperature values of 0.0, 0.4, 0.7, and 1.0 using 500 manually annotated posts, yielding accuracies of 66.32% ± 2.01%, 66.32% ± 1.94%, 66.36% ± 2.14%, and 66.40% ± 2.19%, respectively. As we did not find a significant difference in accuracy by varying the temperature parameter, we adopted 0.4 for the experiments. While the majority of our dataset is labeled using the Llama model, this manual annotation of 500 samples serves as a quality check, providing a measure of confidence in the model's performance.

Table 3.2: Per-class and global average performance metrics for stance classification obtained from 5 independent runs of Llama 3.3 model on 500 (manually annotated) posts, using a temperature of 0.4. Precision, recall, and F1-scores are reported for each class label. Accuracy, "macro avg" and "weighted avg" are reported globally for the set of 500 posts. "Macro avg" indicates the unweighted mean of the per-class metrics, while "weighted avg" accounts for class imbalance. Accuracy is used as a proxy for the agreement between Llama 3.3 annotations and human annotations on 500 sample posts.

Class	**Precision**	**Recall**	**F1-score**	**Support**
against	0.5464 ± 0.0299	0.9282 ± 0.0144	0.6875 ± 0.0241	141
favor	0.7458 ± 0.0187	0.8775 ± 0.0284	0.8062 ± 0.0211	192
neutral	0.8087 ± 0.0307	0.2254 ± 0.0211	0.3520 ± 0.0258	167
accuracy	–	–	0.6632 ± 0.0194	500
macro avg	0.7003 ± 0.0220	0.6770 ± 0.0058	0.6152 ± 0.0133	500
weighted avg	0.7158 ± 0.0179	0.6632 ± 0.0217	0.6170 ± 0.0214	500

Table 3.3 shows some example posts together with the annotations by the Llama 3.3 model by comparison with human annotations. Some of the instances where the Llama 3.3 model made mistakes were posts with aggressive and negative words, or posts where the stance is not explicitly stated and rather subtle, as seen in the last two posts in Table 3.3.

Since the classifier's accuracy (66.32%) at post-level shows moderate agreement with human judgment on a sample of 500 posts and is above the accuracy of the random baseline (33.33%), it can be used to approximate directional changes, specifically, whether the proportion of posts against vaccines has increased or decreased at a monthly aggregated level. This allows us to approximately identify broad trends in vaccine attitudes across the population over the years. However, these estimates should not be treated as precise proportions

Table 3.3: Examples of posts labeled by Llama 3.3 model (llama_label) and human annotators (human_label)

Posts	**llama_label**	**human_label**
The autism one really gets me ... considering it's a genetic defect that you're born with and not something transmitted through vaccines.	in-favor	in-favor
Parents: heard the new vaccine requirement? Meningitis vaccine required for IL students entering 6th, 12th grade.	neither	neither
Then it's not recommended. the govt is making the public suffer by gambling with this, i cant. Make the non-medical govt officials take the vaccine first.	against	against
"One of the last holy grails of HIV research is the development of a HIV vaccine." #homenews #news #cryptonews	neither	in-favor
Watching the age go up for what's considered eligible for the vaccine. KNOCK IT OFF	against	in-favor

or as a representative of the general population.

Figure 3.7 shows the approximate proportion of posts in favor and the proportion against vaccines over the years. We observe that there is a slight increasing trend of the estimated share of posts in favor of vaccines between 2013 and 2020. Starting from 2020 onwards, we see three trends of substantial decrease, substantial increase back to pre-2020 levels, and finally, a substantial decrease in such posts. The drop in 2020 is likely because this phase was marked by posts that talk about the lack of, the need for, and the difficulty of developing a new vaccine for COVID-19. By late 2020, with the news of successful testing of new COVID-19 vaccines, we observe a growth of in-favor posts. However, with the deployment of vaccines in 2021, a general decrease in the virus virulence, and greater awareness of vaccines, the percentage of in-favor posts has steadily declined.

The percentage of against-vaccine posts also has a slight increasing trend from 2013 onwards, which lasts until about 2020. There is, however, a notable decline in such posts in late 2020, and from 2021 onwards, we observe a sharp increase in the percentage of against posts. Notably, by late 2022, our LLM-based estimates suggest that the percentage of posts against vaccines has surpassed those in favor of vaccines. The sharp increase in anti-vaccine stance after 2021 correlates with the introduction of COVID-19 vaccines, their rapid development, and later concerns about their safety and efficacy. The COVID-19 pandemic created a surge of anxiety, uncertainty, and polarized opinions around vaccination, potentially

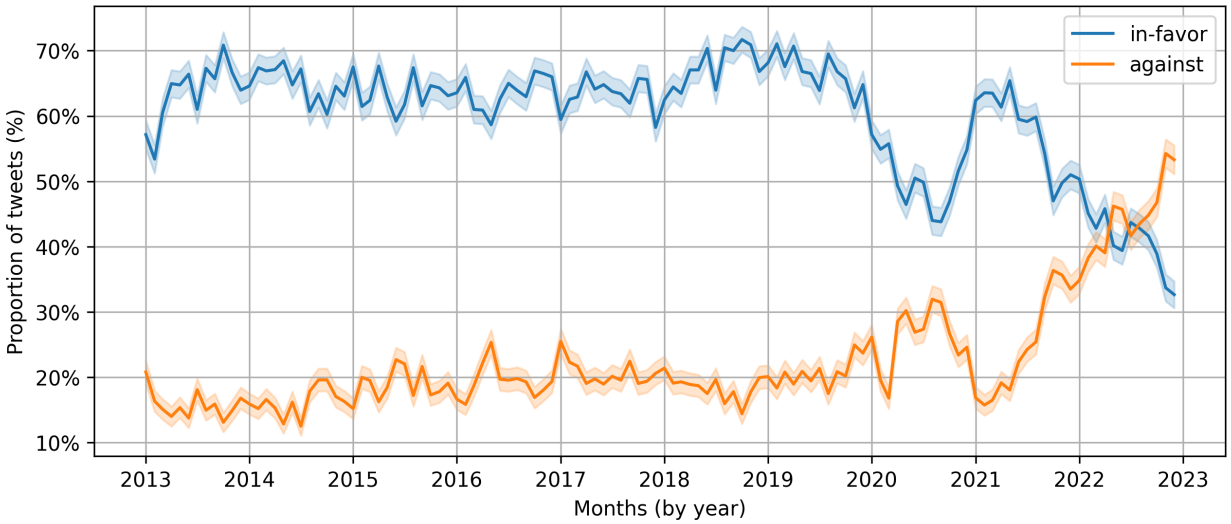


Figure 3.7: Monthly proportion of pro- and anti-vaccine posts from 2013–2022. For each month, 2,000 posts were sampled and classified as in-favor or against vaccination.

leading to a marked decline in pro-vaccine sentiment and a rise in skepticism and opposition.

RQ4. To what extent is vaccine opposition or skepticism characterized by untrustworthiness (low warmth) versus language that questions vaccine effectiveness (low competence)?

Table 3.4 shows the emotion-word density for low warmth and low competence words used per post during pre-COVID-19 (2013–2019) and during COVID-19 (2020–2022) years by English-speaking X users with in-favor and against stances towards vaccines. Emotion-word density is calculated as the total number of emotion-related words divided by the total number of words per post. We observed significant temporal shifts in the emotional tone of vaccine-related discourse from the pre-COVID-19 period to the COVID-19 era. Among posts that were in favor of vaccines, emotion word density associated with low warmth (i.e., untrustworthiness) decreased from 0.1006 to 0.0942, representing a 6.4% decrease (Mann-Whitney U test, $p < 0.001$). Similarly, the density of words associated with low competence emotions declined from 0.0698 to 0.0688, corresponding to a 1.4% decrease (Mann-Whitney U test, $p < 0.001$). In contrast, posts expressing an anti-vaccine stance exhibited an increase in low warmth emotion word density, rising from 0.1199 to 0.1214 (a 1.3% increase; Mann-Whitney U test, $p < 0.001$), while low competence emotion-word density declined from 0.0742 to 0.0725 (a 2.5% decrease; Mann-Whitney U test, $p < 0.001$). These findings suggest

a polarization in emotional framing, with pro-vaccine discourse becoming less negatively charged, whereas anti-vaccine discourse demonstrated a modest intensification of low warmth emotional expression.

Table 3.4: Emotion-word density for low-warmth and low-competence emotions. Emotion-word density is calculated as the proportion of emotion-related words to the total number of words in each post. Percent change expresses the percent increase or decrease from the pre-COVID period to the COVID period, calculated as ((COVID-19 mean - pre-COVID-19 mean) / pre-COVID-19 mean) x 100%.

	Pre-COVID-19 (2013–2019)		**COVID-19 (2020–2022)**		**% change**	
Emotion	In favor	Against	In favor	Against	In favor	Against
Low warmth	0.1006	0.1199	0.0942	0.1214	-6.4%	+1.3%
Low competence	0.0698	0.0743	0.0688	0.0725	-1.4%	-2.5%

To gain more insights into this question, we also performed a tree-map analysis of the top 15 words with low warmth and low competence. The tree-maps for the top 15 words with low warmth for in-favor and against vaccine stances are shown in Figure A1, while the tree maps for the top 15 words with low competence are shown in Figure A2. We find that pro-vaccine stance posts include top words like "die", "bad", "risk", and "shot", focusing on concerns/risks and negative outcomes, along with words like "wear" (referring to mask) and "shot" focusing on preventative actions, whereas anti-vaccine stance posts include top words like "bad", "dangerous", "forced" and "mandatory", suggesting concerns about choice of vaccination. The anti-vaccine stance posts use more emotionally charged negative words, potentially aiming to evoke fear. The pro-vaccine stance, while also using negative terms, focuses more on risks and preventative actions. Similarly, pro-vaccine stance posts use low competence words like "ill", "die", "masks", and "wait", whereas anti-vaccine stance posts use words like "fake", "fear", "injury", and "dead." The anti-vaccine stance posts show more words related to distrust (e.g. "fake", "fear") and lack of personal agency (e.g. "forced", "mandatory"), highlighting their key concerns.

3.5 Conclusions

In this work, we applied the social cognition theory to study the English-language vaccine discourse on X (formerly Twitter) over the past decade, specifically before and during the COVID-19 pandemic. Our analysis demonstrates that vaccine discourse became more emotionally charged during the pandemic, with decreases in negative emotion expressions. The early discourse during the COVID-19 pandemic reflected optimism and trust in vaccine development. However, this was later followed by polarization towards the end of the COVID-19 years. These observations pertain to English-language posts on X and are viewed as corpus-level patterns. Furthermore, our study highlights the potential of utilizing a large-scale, high-quality dataset to examine public vaccine discourse. Spanning over a decade, this dataset offers rich opportunities for future research by leveraging cutting-edge LLM-based techniques, such as self-supervised, semi-supervised, and active learning–to better understand public sentiment and improve the detection of stance and emotion in public health communication.

3.6 Limitations and future research

The scope of this study is limited to English posts on X from 2013 to 2022 that mention vaccines. Thus, the conclusions apply to this dataset and should not, on their own, be used to draw conclusions about people at large. It is known, for example, that X users tend to be younger and technologically savvy compared to the overall population. Most of the X data is not geo-tagged, so we cannot determine the extent to which these posts come from various world regions. Furthermore, the analysis focuses on English-language posts, which may not reflect global trends or cultural differences in vaccine perceptions. Lastly, while the study covers a significant timeframe, it may not capture long-term trends beyond the scope of the data collection period which was limited to a decade between January 1st, 2013 and December 31st, 2022.

That said, since X is such an influential platform, it is useful to better understand the

discourse on vaccines. Future research could address these limitations by incorporating data from multiple social media platforms, improving stance detection results, analyzing non-English content, and extending the timeframe of the study.

As shown in this research, LLMs are able to accurately determine the stance of people towards vaccines in a majority of cases. However, it should be noted that people use language in subtle and nuanced ways (including sarcasm and humor); social media posts do not always provide all the necessary context to understand individual posts; and there is considerable person-to-person difference in how we use language. Thus, while the use of LLMs to determine broad trends in vaccine stance is reasonable, they should not be used on their own to draw conclusions about individual posters.

Our analysis is focused exclusively on textual discourse and does not incorporate social media engagement metrics such as likes, comments, and re-posts. Although these signals could provide insights into content diffusion, influence, and reach, they fall outside the scope of this study. By focusing exclusively on linguistic content, we ensure that our findings are directly tied to the framing and perception of vaccines within public English-language discourse on X. However, this omission constrains our ability to draw conclusions about how widely particular narratives spread or how they shape broader audience engagement patterns.

Using SEC data for monthly trend analysis requires interpolation to estimate monthly trends because Twitter only reported user statistics quarterly (every three months). Therefore, such interpolation may not accurately reflect actual month-to-month fluctuations, seasonal patterns, or event-driven user behavior changes. Additionally, the methodological breakdown between MAU (2013-2019) and mDAU (2019-2022) reporting creates a fundamental discontinuity that prevents direct comparison of user growth trends across the platform's most significant transition period.

While our primary focus is on understanding the shifts in vaccine discourse before and during the COVID-19 pandemic, an important question still remains: Will vaccine discourse return to its pre-pandemic state, or has the landscape permanently shifted? Due to limitations in data collection under the new policy of X, our dataset only extends through the

end of 2022. Examining post-pandemic trends would provide valuable insights into the long-term impact of COVID-19 on public perceptions of vaccines. Future research could explore whether the emotionally charged discourse during the pandemic persists or if the discourse has stabilized over time.

Chapter 4

Evaluating large language models for stance detection on financial targets from SEC filing reports and earnings call transcripts

4.1 Introduction

Financial narratives from U.S. Securities and Exchange Commission (SEC) filing reports and quarterly earnings call transcripts (ECT) constitute the most fundamental sources of information for a wide array of stakeholders. These documents provide comprehensive and fine-grained accounts of a company's financial health, operational performance, strategic initiatives, potential risks, past performance, and management's outlook on future prospects (Bozanic et al., 2017; Hassan et al., 2024). They are highly valuable to investors, auditors, and regulators as they offer critical information about management's perspective on key financial metrics such as debt, earnings per share (EPS), and sales. Despite their importance, these documents are often lengthy and complex, with convoluted sentence structures and use of specialized financial and legal terminologies, making manual analysis both

time-consuming and labor-intensive.

Early work on sentiment analysis on financial texts relied on lexicon-based approaches such as a lexicon dictionary from Loughran and McDonald (2011) and traditional machine learning techniques (Antweiler and Frank, 2004; Chiong et al., 2018; Kogan et al., 2009; Koukaras et al., 2022; Schumaker and Chen, 2009) using feature representation like bag-of-words or Term Frequency-Inverse Document Frequency (TF-IDF) (Kogan et al., 2009; Soong and Tan, 2021). While simple, such techniques struggled with the domain-specific language and contextual nuances of financial texts (Kearney and Liu, 2014; Loughran and McDonald, 2011). Moreover, they classify whether the language is optimistic (positive sentiment) or pessimistic (negative sentiment), but often fail to capture how sentiment varies across different targets within the text. For example, an increase in debt might be seen as either a strategic opportunity or a major risk, depending on the context.

While deep learning techniques like Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) networks improved by modeling sequential and contextual information (Kraus and Feuerriegel, 2017; Mamillapalli et al., 2024; Sohangir et al., 2018), the significant improvement came with the models like Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019; Karanikola et al., 2023; Soong and Tan, 2021) and its domain-specific adaptation on financial texts such as FinBERT (Araci, 2019; Liu et al., 2021b), which captures complex semantic relationships through large-scale pre-training. However, these models still struggle to provide a stance with respect to individual financial targets (Liu et al., 2021b). Additionally, they require thousands of labeled examples per class, which can be very expensive to gather and annotate.

The recent emergence of advanced Large Language Models (LLMs) provides a promising opportunity in the field of NLP and stance detection. With in-context learning, a single prompt can deliver near-state-of-the-art results on unseen stance datasets (Pangtey et al., 2025; Wang and Brorsson, 2025). The ability of these models to perform robustly with minimal or no task-specific labeled data provides significant importance for practical financial applications, given the high cost and effort required to create a large labeled dataset. While LLMs have been used in financial sentiment detection tasks, the research still lacks a

systematic investigation into the precise task of detecting stances towards specific financial targets (debt, EPS, sales) at the sentence level.

In this work, we construct a sentence-level corpus derived from "Form 10-K (Annual Reports)" filed with the U.S. Securities and Exchange Commission (SEC) and quarterly ECTs. The corpus consists of sentences that explicitly reference three financial targets–debt, EPS, and sales–or are relevant to these targets. We annotate these sentences with ChatGPT-o3-pro, an advanced reasoning model from OpenAI, with strict quality control with human validation. Utilizing this dataset, we perform a systematic comparison of various LLMs (Llama3.3, Gemma3, Mistral 3, and GPT-4.1-Mini) under zero-shot, few-shot, and chain-of-thought prompting scenarios. This approach is important for assessing the potential of LLMs in real-world financial analysis scenarios where the availability of extensive, target-specific annotated data is often limited.

We make our prompts, data, and code publicly available for reproducibility and advanced research on LLM-based financial stance detection.

4.2 Related work

Research on sentiment analysis and stance detection in financial texts has progressed from lexicon-based models and traditional machine learning models to transformers and large language models. Loughran and McDonald (2011) showed that general negative word dictionaries misclassify common financial terms and proposed a domain-tailored lexicon that better captures positive/negative tone in 10-K reports, linking those sentiment measures to stock market reactions (e.g., returns and volatility). Mukherjee et al. (2022) introduced ECTSum, a dataset curated by summarizing long ECTs. Various works utilized traditional machine learning techniques including regression models (Kogan et al., 2009; Koukaras et al., 2022), Support Vector Machines (SVMs) (Chiong et al., 2018; Schumaker and Chen, 2009), Naive Bayes classifiers (Antweiler and Frank, 2004), and tree-based methods like Random Forests (Koukaras et al., 2022) that typically use feature representation like bag-of-words or Term Frequency-Inverse Document Frequency (TF-IDF) (Kogan et al., 2009; Soong and Tan,

2021). Similarly, deep learning techniques like Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) networks have been used in sentiment analysis in the financial domain (Kraus and Feuerriegel, 2017; Mamillapalli et al., 2024; Sohangir et al., 2018). More recent works utilize Bidirectional Encoder Representations from Transformers (BERT) Devlin et al. (2019) and its domain-specific adaptation on financial texts such as FinBERT (Araci, 2019; Cicekyurt and Bakal, 2025; Karanikola et al., 2023; Liu et al., 2021b; Peng et al., 2021). Singh et al. (2023) introduced Fin-STance, a deep learning-based multi-task model specifically designed for detecting both financial stance and sentiment from financial data.

The most recent LLMs trained with a massive volume of texts show superior natural language understanding and outperform smaller transformer-based models like FinBERT (Kang and Choi, 2025). Various works have looked into utilizing the power of such LLMs in sentiment analysis and research in the financial domain (Fatouros et al., 2023; Feng et al., 2025; Guo et al., 2023; Huang et al., 2024; Li et al., 2023a; Wei and Liu, 2025; Zhang et al., 2023). Wang and Brorsson (2025) studies various zero-shot, few-shot, and fine-tuning-based approaches with LLMs for financial text analysis. A comprehensive review of methods for sentiment (stance) analysis in financial texts is conducted by Du et al. (2024, 2025). In addition, Nie et al. (2024) provides an extensive survey of the applications of LLMs in the financial domain.

4.3 Financial dataset

4.3.1 SEC form 10-K annual report

The U.S. SEC Form 10-K is a comprehensive annual report mandated for publicly traded companies that provides a detailed overview of financial performance, operational activities, and corporate governance (Cazier and Pfeiffer, 2016). These reports are essential resources for investors, analysts, and regulatory bodies as they provide comprehensive insights into corporate strategic initiatives and identified risk factors. Specifically, "*Section 7, Manage-*

Table 4.1: Examples of instances in the ECT and the SEC dataset for each target and stance class.

ECT examples	SEC examples	Target	Stance
as a result of the strong EBITDA growth and free cash flow, we're on a path to see adjusted net leverage drop to approximately 3x by the end of the year, absent any other actions.	*Total debt decreased $6.4 million to $194.4 million at December 31, 2020 from $200.8 million at December 31, 2019.*	debt	Positive
we estimate at least $0.50 of adjusted eps accretion next year as the business returns toward pre-covid levels.	*Weighted-average diluted shares outstanding in 2019 declined 2.8 percent year-on-year which benefited earnings per share.*	EPS	Positive
the strong sales mix also led to excellent profitability.	*In addition, selling prices increased year-on-year by 0.6 percent for full-year 2020, and lower raw-material costs reduced cost of sales as a percentage of sales.*	sales	Positive
adjusted net leverage was 3.7 times at year-end, up slightly from Q3 reflecting the impact of the input-cost increases on EBITDA.	*We cannot assure that our operating performance, cash flow and capital resources will be sufficient to repay our debt in the future.*	debt	Negative
when combined, we estimate these two factors had more than a $0.20 negative impact on adjusted eps in the quarter.	*Divestiture impacts include the lost operating income from divested businesses, which decreased earnings per diluted share by 3 cents year-on-year for 2020.*	EPS	Negative
so, we announced a 10% decrease in the business for the segment, but that was really driven by two major factors.	*Total sales decreased 14 percent in Mexico, which included decreased organic sales of 12 percent.*	sales	Negative
net debt finished the year at just under $1.2 billion.	*As of December 31, 2019, we had approximately $21.6 million of LIBOR-based debt.*	debt	Neutral
on this slide, you can see the components that impacted our operating margins and earnings per share performance as compared to Q1 last year.	*A discussion related to the components of year-on-year changes in operating-income margin and earnings per diluted share follows: Organic growth/productivity and other.*	EPS	Neutral
just can you sustain that sales momentum?	*Sales grew in home improvement and home care, while consumer health care and stationery and office declined.*	sales	Neutral

ment's Discussion and Analysis (MD&A)" serves as a primary focus for textual analysis, as it encompasses management's interpretative narrative regarding financial outcomes, examination of recognized uncertainties, and forward-looking statements about prospective developments (Amel-Zadeh and Faasse, 2016).

4.3.2 Quarterly earnings call transcripts

Quarterly ECTs are a textual record of teleconferences held by a company's management with financial analysts, investors, and the media, usually following the release of quarterly financial results (Mukherjee et al., 2022). These transcripts serve as a valuable source of

Table 4.2: Distribution of Positive, Negative, and Neutral stance labels for each financial target considered (debt, EPS, and sales) in SEC filing reports (SEC) and earnings call transcripts (ECT), shown separately for the training and test splits. Pos., Neg., and Neu. represent Positive, Negative, and Neutral, respectively.

		Train				**Test**			
Dataset	**Target**	Pos.	Neg.	Neu.	Total	Pos.	Neg.	Neu.	Total
SEC	debt	27	50	8	85	50	35	12	97
	EPS	10	32	3	45	14	14	9	37
	sales	70	72	18	160	59	100	12	171
ECT	debt	73	7	10	90	65	12	20	97
	EPS	44	37	16	97	84	65	22	171
	sales	79	52	14	145	129	127	39	295

qualitative data by capturing the dialogues between senior executives and financial analysts and offer a more interactive and dynamic medium compared to static documents, such as SEC filings, allowing executives to explain performance metrics in greater depth, provide necessary context, and address analysts' queries directly (Bushee et al., 2003).

4.3.3 Dataset collection

We utilize SEC filing reports from two companies–MATIV Holdings Inc. and 3M Co.–using data from 2020-2021 as the training set and data from 2022-2024 as the test set. MATIV Holdings Inc. and 3M Co. operate in the same industry but have hugely different market caps. Similarly, we also use ECTs from these two companies from 2020-2021 as training data and 2022-2024 as a test. We transformed the text from SEC reports and ECTs into individual sentences using PyPDFLoader[1] from the LangChain library. We then used the Llama-3 model to remove irrelevant sentences, keeping only those sentences containing key phrases related to financial targets of interest, including debt, EPS, and sales. Appendix C.1 shows the prompt used for filtering relevant sentences. Our initial results indicated that LLaMa-3 generated a significant number of false positives, incorrectly classifying irrelevant

[1] https://python.langchain.com/docs/integrations/document_loaders/pypdfloader/

sentences as relevant. To address this, we applied the same prompt to ChatGPT, whose superior reasoning abilities allowed us to filter out most of these false positives. We then annotated the relevant sentences using ChatGPT-o3-pro, an LLM with advanced reasoning, deep analytical thinking, and complex problem-solving capabilities (OpenAI, 2025). The model generated both a stance label and a corresponding justification for each label.

We validated ChatGPT-o3-pro's annotations and justifications of ECT sentences for MATIV Holdings Inc. by having human annotators evaluate their correctness. Annotators reviewed the stance label and accompanying justification produced by the model, indicating whether they agree with its reasoning. Across all evaluated targets—sales, EPS, and debt—human annotators showed over 97% agreement with the justifications provided by ChatGPT-o3-pro, indicating a high level of alignment between the model's analytical output and human judgment. Given this high agreement, we utilized ChatGPT annotations as ground truth for the ECTs of 3M Co. and SEC report statements from both companies (see Table A1 for samples for each target and stance class from two datasets). Table 4.2 shows the statistics of our final dataset. We have, on average, 122 training and 107 test instances per target in the SEC dataset, whereas the ECT dataset has, on average, 120 training and 193 test instances per target.

4.4 Methodology

4.4.1 Large language models

For a comprehensive comparative analysis, we employ four large language models: three open-source or open-weight models–Llama 3.3, Gemma3-27B, and Mistral 3 Small–and one proprietary model, ChatGPT 4.1-Mini. A brief overview of each model is provided below:

1. **Meta LLaMa 3.3:** The Llama 3.3 model (Llama3.3:70B) from Meta is a 70 billion-parameter instruction-tuned, text-only generative model specifically developed for instruction-following tasks. It is optimized for multilingual dialogue and uses an optimized transformer architecture (Grattafiori et al., 2024). The tuned version uses Supervised Fine-

Tuning (SFT) and Reinforcement Learning with Human Feedback (RLHF) to align with human preferences for helpfulness and safety. The model supports a context window of up to 128,000 tokens and is optimized for inference efficiency through Grouped-Query Attention. In evaluations, it outperforms both its predecessor (Llama 3.1-70B) and larger models like Llama 3.1-405B. Consequently, Llama 3.3 provides a robust and efficient option for research and production use in conversational AI.

2. **Gemma-3-27B:** Gemma 3 (Gemma3:27B) is a 27 billion-parameter instruction-tuned model from Google DeepMind (Kamath et al., 2025). This model employs a multimodal architecture with vision understanding abilities, extends multilingual coverage, and has a longer context of at least 128,000 tokens. Using distillation-based pre-training and a novel post-training alignment phase, Gemma 3 significantly improves the math, chat, and instruction-following abilities, making its capabilities comparable to the more advanced and larger Gemini-1.5-pro model (Kamath et al., 2025).

3. **Mistral 3 small:** The Mistral Small 3 (Mistral3:24B) is a 24-billion-parameter, instruction-tuned decoder-only transformer model optimized for low latency token generation (MistralAI, 2025) and has a context window of 32,000 tokens. The combination of strong multilingual coverage, competitive reasoning ability, and high throughput makes Mistral Small 3 a compelling open-weight foundation model for research and production-grade conversational systems that requires low latency and modest hardware (MistralAI, 2025).

4. **ChatGPT 4.1-Mini:** The ChatGPT 4.1-Mini is OpenAI's "mini" variant of the proprietary GPT-4.1 family (OpenAI, 2025). We utilized the *gpt-4.1-mini-2025-04-14* model snapshot for our performance comparison. This model has a very large context window of 1 million with full multimodal (text + image) input support. The GPT-4.1 mini is a fast and efficient small model, delivering significant improvements compared to the GPT-4o mini in instruction-following, coding, and overall intelligence (OpenAI, 2025).

LLaMa 3.3, Gemma-3-27B, and Mistral 3 Small models are instruction-tuned, transformer-based LLMs with sufficiently large context windows. GPT 4.1-Mini, a proprietary model, has its exact parameter count undisclosed. OpenAI confirms it is considerably lighter than the full GPT-4.1 model (OpenAI, 2025). Because its scale is roughly comparable, we selected GPT 4.1-Mini as a benchmark when comparing the two open-weight alternatives–Llama 3.3-70B.

4.4.2 Experimental setup

In our study, we experiment with LLMs on two primary datasets: the SEC reports and the ECTs. For both datasets, we study the usefulness of providing relevant background information about the company as *context* for the models in the prompt. Specifically, we extract "*Section 7 Management's Discussion and Analysis of Financial Condition*" from the SEC report and the complete transcript from quarterly earnings calls, respectively. We experimented with three scenarios–zero-shot, few-shot, and context usage from the transcripts–each with and without chain-of-thought (CoT) reasoning prompts or demonstrations.

Context usage scenarios: We evaluate the impact of providing background information about the company as additional context. We specifically investigate three scenarios:

1. *No context:* In this scenario, neither Section 7 from the SEC report nor the ECTs are included in the prompt as additional context to the LLMs. This serves as a control for our experiment to study how adding additional context about the company affects the model's performance.

2. *Full context:* In this scenario, we provide the entire "*Section 7, Management's Discussion and Analysis (MD&A)*" content from the SEC report or the complete ECT as additional context to the LLMs in their corresponding prompts for SEC and ECT, respectively.

3. *Summarized context:* In this scenario, given that "*Section 7, Management's Discussion and Analysis (MD&A)*" of the annual SEC reports spans over 20 pages on average, we

summarize the Section 7 content from the SEC report using ChatGPT-o3-pro model to create more concise background information. Similarly, we also summarize the entire ECT using the ChatGPT-o3-pro model to study the differences in the performance of LLMs using summarized content.

Prompting scenarios: We evaluate three prompting scenarios:

1. **Zero-shot setting:** In the zero-shot scenario, we evaluate how well LLMs perform without any labeled examples. We examine the performance across the three context conditions described above: (1) zero-shot with no context, (2) zero-shot with full context, and (3) zero-shot with the summarized context. This enables us to assess how the background information affects the model's ability to generalize without exposure to task-specific examples.

2. **Few-shot setting:** In the few-shot setting, we include a small number of labeled examples from the training set in the prompt to guide the LLM's predictions. We investigate two strategies for few-shot example selection: (1) random sampling and (2) selecting examples from the training set that are most semantically similar to the test instance. Similar examples are selected based on the highest cosine similarity between the test instance and the examples from the training set. The embeddings to calculate the cosine similarity are generated by using Sentence-BERT (Reimers and Gurevych, 2019). We explore the effect of varying the number of examples, k, per stance class ($k = 1, 5, 10$) for each target and dataset. Similar to the zero-shot setting, we also test three background information configurations: (1) a few-shot with no context, (2) a few-shot with full context, and (3) a few-shot with summarized context.

3. **Chain-of-Thought setting:** In the Chain-of-Thought (CoT) setting, we prompt the LLMs to generate intermediate reasoning steps before arriving at a final prediction. This approach is particularly important for tasks that require multi-step reasoning. We evaluate the CoT performance using the same three context usage scenarios: (1) No context, (2) Full context, and (3) Summarized context. We also evaluate CoT

Table 4.3: Few-shot classification accuracy (%) on the ECT and SEC datasets. Values are mean±std over three runs; for every model–dataset pair, the higher score between *Random* and *Most similar* sampling is **highlighted**.

	ECT		SEC	
Model	**Random**	**Most similar**	**Random**	**Most similar**
GPT-4.1-Mini	86.73±0.52	**87.60±0.75**	83.65±1.22	**85.29±1.20**
Gemma3:27B	85.63±0.53	**86.43±0.45**	76.55±0.87	**79.21±0.96**
Llama3.3:70B	84.36±1.76	**87.04±2.07**	79.14±2.10	**79.42±2.35**
Mistral:24B	76.49±2.10	**80.11±2.03**	62.69±3.01	**67.07±2.58**

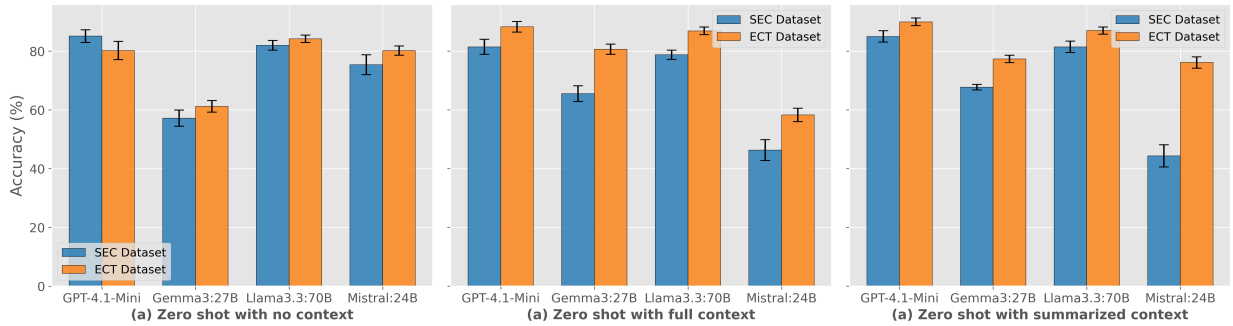
performance in zero-shot and few-shot settings. This allows us to investigate how explicit reasoning steps combined with varying levels of contextual input and class-specific examples impact model performance and generalization.

4.5 Results and discussion

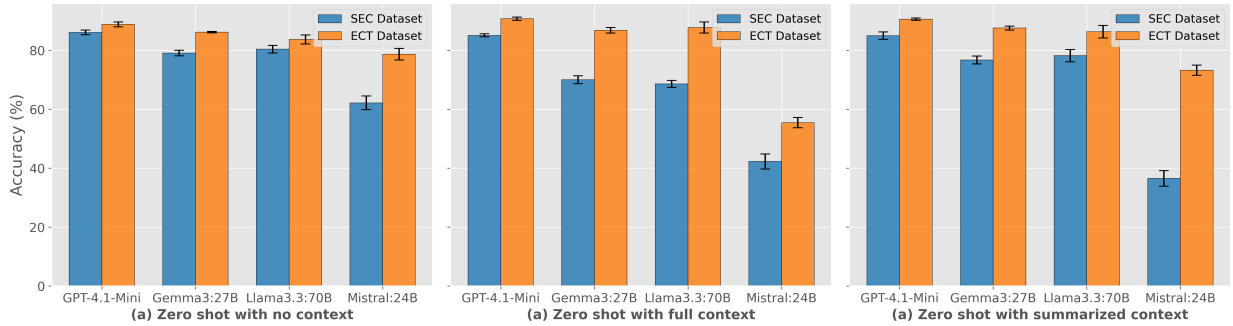
The numeric results for both datasets, encompassing all models and experimental configurations, are summarized in Appendix Table C.6. Our experimental results show that overall, ChatGPT-4.1-Mini achieves the highest performance with an average accuracy across all experimental setups of 87.79% ± 1.47, followed by Llama3.3:70B (83.02% ± 1.98). Gemma3:27B shows comparable performance with an accuracy of 81.21% ± 1.48 while Mistral:24B performs the worst with 68.6% ± 2.71 accuracy. In what follows, we discuss the detailed findings from our experiments across various experimental setups.

4.5.1 Context usage: no context, full context, and summarized context

Figure 4.2 shows the few-shot performance of the models on the ECT dataset (top row) and the SEC dataset (bottom row) across three context usage scenarios. Within each figure, accuracy (mean ± std) is plotted against the number of few-shot examples $k = 0, 1, 5, 10$



(A) **With chain-of-thought prompts.**



(B) **Without chain-of-thought prompts.**

Figure 4.1: Zero-shot accuracy of models on the SEC and ECT datasets (A) *with* and (B) *without* chain-of-thought (CoT) prompting. Each of these models and CoT settings is evaluated across three transcript-usage scenarios: (a) no transcript context, (b) full transcript context, and (c) summarized context.

for three context usage scenarios: No context (red), Full context (blue), and Summarized context (light purple). We find that incorporating contextual information–either full or summarized– markedly improves the zero-shot performance on the ECT dataset for GPT-4.1-Mini and Gemma3:27B, whereas there is no significant improvement for the Llama3.3:70B model, and performance even degrades for the Mistral:24B model. For the SEC dataset, only Gemma3:27B model shows significant performance improvement using the context, while the performance of Mistral:24B shows a significant drop in the zero-shot setting using the context. For models that benefit from additional context, we find that models perform similarly in both full and summarized context usage, suggesting that the summarized context is nearly as effective as the full context.

As the number of few-shot examples increases, we find that all model performs similarly

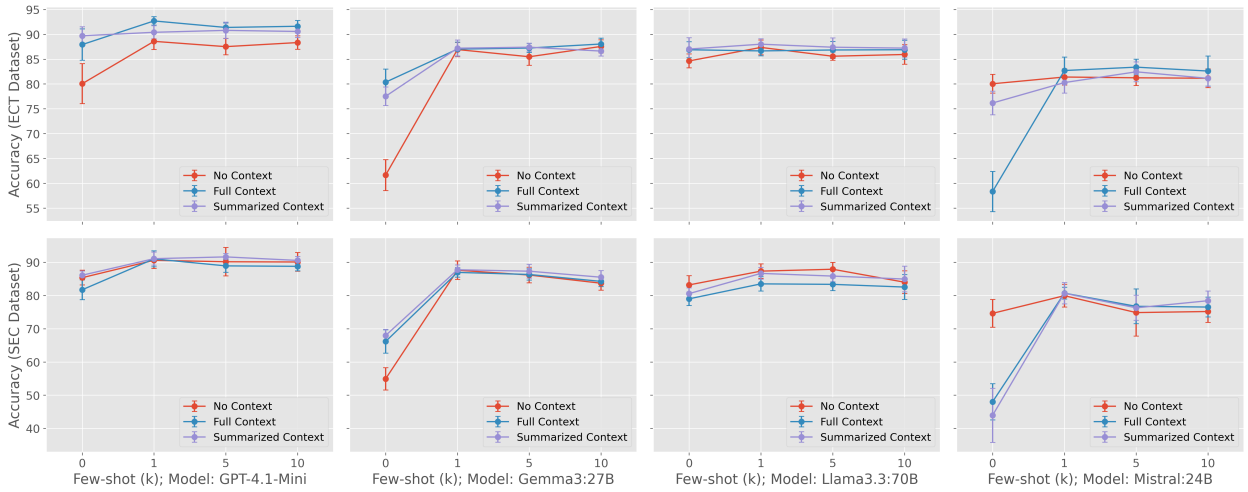


Figure 4.2: Context usage scenarios across different models on two datasets. Few-shot classification accuracy on the ECT dataset (top row) and the SEC dataset (bottom row) using three context usage scenarios: a) No Context, b) Full context, and c) Summarized Context across four models. Columns, left to right, represent GPT-4.1-Mini, Gemma3-27B, Llama3-70B, and Mistral-24B. k represents the number of most similar examples with a chain-of-thought demonstration per class. Error bars represent the standard deviation over three independent runs.

regardless of the context provided (no, full, or summarized context) for both datasets. In this scenario, contextual information provides minimal additional benefit, indicating that models prioritize in-context learning from the provided few-shot examples over utilizing the contextual information. This finding suggests that when a sufficient few-shot examples are available, extensive context becomes less important for achieving a good performance.

4.5.2 Random vs. semantically similar few-shot examples

Table 4.3 shows the performance comparison of various LLMs when using randomly selected examples versus semantically most similar few-shot examples across both ECT and SEC datasets. Across all models and both datasets, selecting the most similar examples consistently improves the accuracy over randomly selecting the few-shot examples. For the ECT dataset, the average performance improvement across all models when using semantically similar examples is 2%, whereas for the SEC data, the average improvement is slightly higher, 2.24%. The most notable performance improvement is observed with the Mistral:24B, with

an average accuracy improvement of 4% across both datasets. The remaining models show modest performance improvement with average improvement of 1.73% for Gemma3:27B, 1.48% for Llama3.3:70B, and 1.26% for GPT-4.1-Mini.

Our experimental results show that selecting the semantically most similar few-shot examples consistently yields superior performance compared to selecting few-shot examples randomly. These findings indicate that carefully selecting examples that are semantically similar to the test instance provides more relevant context and clearer guidance for LLMs, thereby improving stance detection across diverse linguistic scenarios.

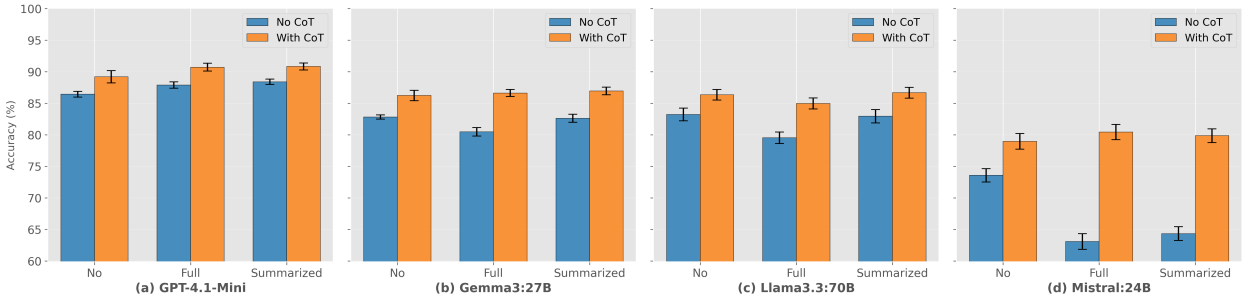


Figure 4.3: Few-shot performance of transcript usage scenarios *with* and *without* chain-of-thought (CoT) prompts. Accuracy is shown for four models–(a) GPT-4.1-Mini, (b) Llama3.3:70B, (c) Gemma3:27B, and (d) Mistral:24B. The result is averaged for both SEC and ECT data.

4.5.3 Effectiveness of Chain-of-Thought (CoT) prompting

Figure 4.3 shows the few-shot performance of various models, comparing scenarios with and without chain-of-thought (CoT) prompting demonstration across various transcript usage scenarios (see Appendix C.4 for an example). The results are averaged over both ECT and SEC datasets. We find that incorporating CoT reasoning in the few-shot examples consistently improved performance across all evaluated LLMs compared to the scenario without CoT demonstrations. The accuracy improvement with CoT demonstrations is 4.23% ± 0.47 averaged across all models and transcript usage scenarios. The most significant improvement is observed for the Mistral:24B model, particularly when using Full and Summarized transcripts. Llama3.3:70B and Gemma3:27B show moderate improvement, while GPT-4.1-Mini

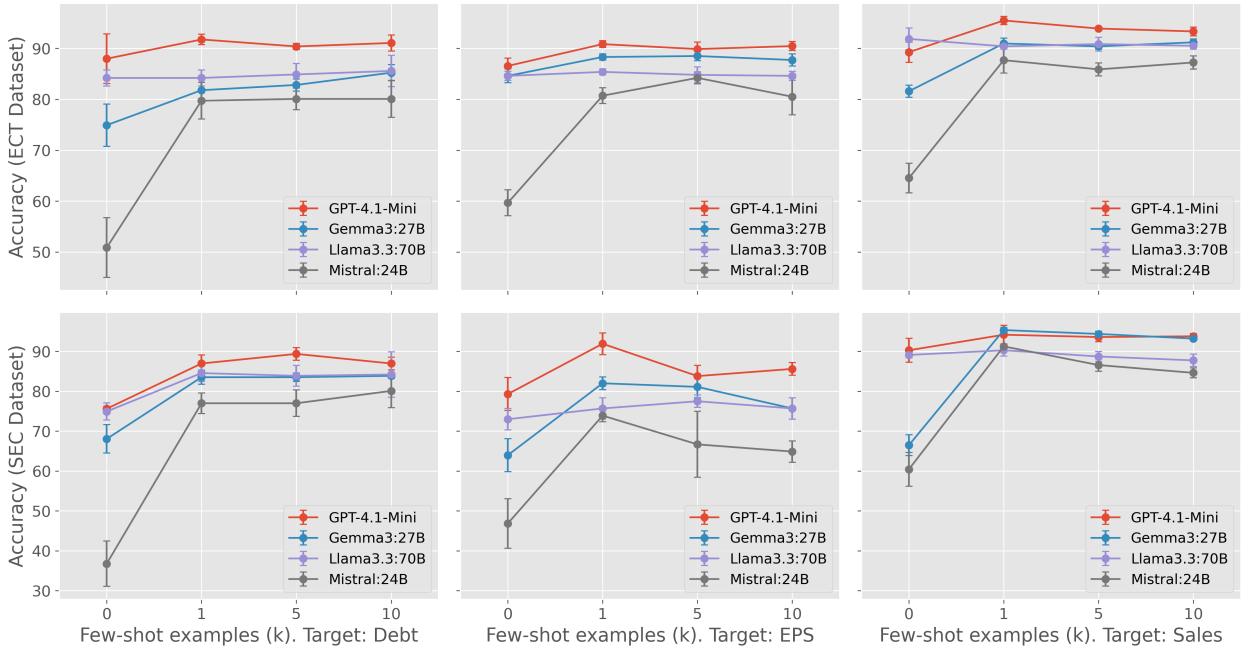


Figure 4.4: Few-shot with chain-of-thought accuracy on two datasets across various targets. Few-shot classification accuracy on the ECT dataset (top row) and the SEC dataset (bottom row) using chain-of-thought prompting for three targets—debt (left), EPS (centre), and sales (right). k represents the number of most similar examples with a chain-of-thought demonstration per class. Error bars represent the standard deviation over three independent runs.

shows the smallest improvement.

These findings suggest that CoT prompting explicitly encourages structured reasoning, enhancing the models' reasoning capabilities (Wei et al., 2022b), which improves the performance in the stance detection task. This CoT prompting is especially beneficial for relatively smaller models, which may rely more on guided reasoning to bridge gaps in implicit understanding of text (Ranaldi and Freitas, 2024).

4.5.4 Few-shot prompting vs. zero-shot prompting

Figure 4.4 shows the performance of various LLMs under zero-shot and few-shot prompting strategies, using $k = 1, 5, 10$ examples with CoT demonstration, across three targets–debt, eps, and sales–for both ECT and SEC datasets. We find that, on average, few-shot prompting with a single example $k = 1$ consistently yields better performance over the zero-shot setting

across all targets and datasets. The improvement is significant for the Mistral:24B model. All models except Llama3:70B show consistent improvement over zero-shot across all targets for $k = 1, 5, 10$ examples in both datasets, suggesting the robustness of few-shot prompting in enhancing models' generalization and in-context learning (Brown et al., 2020; Kojima et al., 2022). Interestingly, the most significant performance improvement is seen with $k = 1$, implying that even a minimal example with CoT guidance can be effective. With $k = 5, 10$ we do not see significant improvement, indicating that additional few-shot examples have diminishing returns as adding more examples increases the prompt length and increases the chance of noisy or conflicting chain-of-thoughts, thus decreasing accuracy (Levy et al., 2024; Liu et al., 2024). These findings highlight the general importance of few-shot learning for stance detection. They also show that model-specific variations exist, suggesting the importance of tuning the optimal number of examples,k, for different LLM architectures.

4.5.5 Performance across SEC and ECT datasets

Figures 4.1A and 4.1B show the performance of the LLMs on the SEC and ECT datasets under identical experimental conditions–specifically, with and without chain-of-thought prompts and varying usage of transcript contexts:(a) No context, (b) Full context, and (c) Summarized context. Similarly, Figure 4.4 shows the performance of the LLMs on the ECT and SEC datasets using few-shot examples ($k = 1, 5, 10$) with chain-of-thought (CoT) prompting across all three targets: debt, eps, and sales.

We find a consistent trend where models achieve similar higher accuracy on the ECT dataset as compared to the SEC dataset across all similar evaluated conditions. Similarly, across experiments with few-shot prompting with $k = 1, 5, 10$ examples, all models consistently achieve higher accuracy on the ECT data than on the SEC data. This performance gap is attributed to fundamental differences in linguistic styles between these two datasets. ECT data consists of conversational statements and often narrates explicit financial changes (e.g., increases or decreases in debt, eps, or sales), providing clearer stance indicators. In contrast, the SEC dataset comprises formally structured sentences that are often numeri-

cally dense and embed key financial indicators in complex syntactic and semantic structures. Consequently, identifying the stance within the SEC dataset requires deeper quantitative reasoning capabilities, which continue to pose significant challenges for current LLMs.

For example, in a SEC instance–"*Our total debt to capital ratios, as calculated under the amended Credit Agreement, at December 31, 2022 and December 31, 2021 were 59.0% and 65.1%, respectively.*", the stance is positive towards debt as the debt-to-capital ratio fell from 61.1% to 59.0%, indicating reduced leverage. However, identifying this requires the model to compute a relative change and understand its implications. Llama3.3:70B incorrectly classifies this as neutral, with justification–*The sentence provides a factual comparison of debt to capital ratios without expressing a clear positive or negative stance towards debt, merely stating the ratios for two years–which lacks understanding of the implication of this change in ratios.* In contrast, an ECT stance–"*Net debt stands at $13.3 billion, up approximately 2% as we continue to invest in the business.*"–has a negative stance towards debt and presents a more direct cue (i.e., an explicit increase in debt), making it easier for the LLMs to correctly identify the negative stance. These examples highlight the greater complexity and reasoning involved in the SEC data, which contributes to the observed accuracy disparities.

4.6 Conclusions

In this study, we systematically evaluated several modern LLMs for stance detection in financial texts at the sentence level, focusing on three financial metrics as targets–debt, EPS, and sales. We introduced a sentence-level stance detection corpus derived from SEC Form 10-K filings and quarterly earnings call transcripts. Using the advanced ChatGPT-o3-pro model, we annotated the sentences from SEC reports and earnings call transcripts with rigorous human validation. We evaluated multiple prompt-based learning strategies, including zero-shot, few-shot, and chain-of-thought (CoT) prompting. Our findings indicate that few-shot prompting enriched with CoT reasoning yields superior performance over zero-shot and without CoT prompting. Particularly, the GPT-4.1-Mini model consistently outperformed other evaluated LLMs, followed by the Llama3.3:70B model.

Our analysis further highlights the importance of the selection of semantically similar few-shot examples in improving the model performance. Furthermore, we observed notable performance differences across two document types–ECT and SEC filing reports. While ECT data provided relatively easier stance detection due to its conversational nature and more explicit financial references, SEC reports posed greater challenges due to their formal complexity and numerical density, emphasizing the need for advanced contextual understanding and reasoning capabilities. Our findings highlight the practical viability of leveraging LLMs for the target-specific stance detection task in the financial domain without requiring extensive labeled data.

4.7 Limitations

Our study has a few limitations that should be acknowledged. First, the dataset is limited in scope as we exclusively focused on two companies–MATIV Holdings Inc. and 3M Co.–which may constrain the generalizability of the findings across broader financial contexts. Second, we rely on the ChatGPT-o3-pro model for dataset annotation. While the agreement with human validation is very high (over 97%), the use of a language model for data annotation may introduce model-induced biases. Specifically, the results generated by ChatGPT-4.1-Mini could be subject to such biases.

Chapter 5

Modular prompt optimization for stance detection

5.1 Introduction

Stance detection, a subfield of opinion mining, is an important task in Natural Language Processing (NLP) that involves determining the attitude or position expressed by the author of a text regarding a particular topic or claim, generally referred as target. The task involves automatically determining if the author of a text is *In-Favor (Positive)*, *Against (Negative)*, or *Neutral* towards a given target (Mohammad et al., 2016). Stance detection has broad applications, ranging from misinformation detection to public opinion mining on political and health issues (ALDayel and Magdy, 2021; Küçük and Can, 2020).

Despite substantial progress in natural language understanding, large language models continue to face limitations in stance detection, particularly in recognizing subtle pragmatic cues such as sarcasm and irony in social media discourse (Gole et al., 2024; Li et al., 2025; Yakura, 2024). Directly fine-tuning or aligning an LLM for stance detection (e.g., via reinforcement learning from human feedback (RLHF)) is typically resource-intensive and requires access to model parameters - an option not available for many closed or API-based systems. Furthermore, LLMs are sensitive to input prompts, few-shot examples, and even the order

of such examples (Lu et al., 2022; Shi et al., 2024; Zhao et al., 2021). Prompts with semantically similar meanings could lead to significantly different performance outcomes (Kojima et al., 2022; Zhou et al., 2022). Additionally, the optimal prompt often depends on both the specific LLM and the particular task at hand (Chen et al., 2023a). Therefore, prompt optimization is often an important step in achieving strong performance with LLMs (Schulhoff et al., 2024). Manual prompt optimization is time-consuming and expensive.

These challenges have prompted systematic approaches to prompt optimization in LLMs, particularly for tasks such as stance detection and related classification problems. Work in this area includes treating prompt design as a structured problem with tunable components (e.g., task description, examples, output format) (Juneja et al., 2024; Khot et al., 2023), framing prompts as programs or graphs optimized through algorithmic search (Schnabel and Neville, 2024), and performing automated prompt refinement where LLMs iteratively improve the prompts (Yang et al., 2024; Zhou et al., 2022). Search and learning strategies such as reinforcement learning and evolutionary algorithms have also been applied (Pryzant et al., 2023; Zeng et al., 2024). Collectively, these methods underscore the importance of automated prompt engineering in many NLP tasks.

In stance detection, recent studies have shown promising results with LLMs by using techniques like chain-of-stance prompts, which decompose the task into reasoning steps (Ma et al., 2024). However, much of this progress stems from manual prompt engineering or reasoning enhancements (Aiyappa et al., 2024; Gül et al., 2024), rather than systematic and automated prompt optimization tailored to unique challenges in stance detection. While automated prompt optimization frameworks have been explored in general text classification tasks (Liu and Shi, 2024), they remain underexplored in the stance detection domain–especially for tasks that require nuanced interpretation of implicit stances and domain-specific contexts.

To address this, we introduce **Mo**dular **Pr**ompt **O**ptimization for **S**tance **D**etection (**MoPrO-SD**), a framework designed specifically for stance detection. Starting with an initial, human-written, base stance detection prompt that consists of a set of distinct modules, we optimize each module independently using an LLM-based optimization process. The prompt modules correspond to specific logical units (e.g., task definition, guidelines, label

descriptions, output format), and and are refined independently through a meta-prompting strategy. In this process, the LLM suggests module-specific improvements based on the current global prompt and performance feedback from the stance detection task. By iterating this process, the framework does prompt revisions in an iterative, modular fashion, rather than altering the entire prompt at once.

We evaluate the proposed framework through a set of comprehensive experiments on a variety of stance detection datasets and LLMs. The datasets used include several established stance detection benchmarks (SemEval-2016, COVID-19-Stance, GunStance), as well as a newly curated VaccineStance dataset consisting of vaccine-related tweets collected over a ten-year period (2013-2022). Together, these datasets cover a variety of domains such as political ideologies, public health and online debates, and help evaluate the generalizability of the proposed (**MoPrO-SD**) framework. To ensure robustness, we conduct experiments with multiple LLMs of different sizes and architectures, including Llama3.3:70B, Gemma3:27B, Phi-4:14B and GPT-oss:20b. We also investigate alternative strategies for optimizing each module (e.g., instruction only, instruction with reasoning), along with the effect of zero-shot and few-shot settings during the prompt refinement. Empirical results show that our modular prompt optimization yields significant improvement in stance detection accuracy, outperforming the baseline human-crafted prompt, across different datasets and different model families. These findings suggest that breaking prompts into modules and optimizing them in a targeted way is an effective paradigm for enhancing performance on the stance detection task.

Our contributions can be summarized as follows:

1. **Modular prompt optimization framework:** We propose a modular prompt optimization framework that can effectively improve the performance of prompt-based stance classification using LLMs.

2. **Extensive evaluation:** We conduct an extensive experimental evaluation of our framework using various LLMs and existing benchmark datasets, as well as a *novel human-annotated VaccineStance dataset*, to demonstrate the effectiveness of the pro-

posed approach.

3. **Empirical insights:** We provide systematic insights into optimization strategies, prompting settings, and the benefits of stronger external reasoners, offering practical guidance for future stance detection research.

5.2 Related work

Prompt engineering has seen a critical focus in terms of adapting and using LLMs for specific downstream tasks, including stance detection, where task-specific guidance is very important, given the LLMs' sensitivity to prompts. In this section, we review recent works related to prompt engineering for stance detection, as well as structured prompting, in-context prompting and automated prompt optimization techniques.

Prompt Engineering for Stance Detection

Recent studies have approached the prompt engineering for LLMs in stance detection from various perspectives. Ma et al. (2024) proposes chain-of-stance, which improves stance classification through intermediate reasoning steps. Lee et al. (2024) highlights the effectiveness of rationale-augmented prompting for smaller models. Gül et al. (2024) fine-tuned LLMs on stance datasets to achieve strong results; however, this approach requires access to model weights. Aiyappa et al. (2024) showed that zero-shot prompting with an instruction-tuned 11B model (Flan-T5-XXL) can approach supervised performance on stance benchmarks, given carefully chosen prompts and decoding strategies.

Structured Prompting and In-context Learning

Some works treat prompt engineering as a structured optimization problem. These include decomposing prompts into multiple loosely coupled semantic sections (Juneja et al., 2024), treating them as structured programs that can be optimized at compile-time (Schnabel and Neville, 2024), and jointly optimizing the format and content of the prompt through

an iterative refinement process (Liu et al., 2025b). Khattab et al. (2023) proposed DSPy, a framework that compiles LLM pipelines into modular instruction-optimizing programs. Along the same lines, some have explored diversity-aware and decision-theoretic optimization frameworks, leveraging the structured selection of prompts to enhance model robustness and decision-making capabilities (Liu et al., 2025a; Pang et al., 2025). Likewise, several studies utilize demonstration-based selection methods to identify influential examples to improve in-context learning (Wan et al., 2025; Wu et al., 2024; Zhu et al., 2024). The mixture-of-demonstrations strategy (Wang et al., 2024a), for instance, selects diverse high-value examples to stabilize LLM predictions.

Automated Prompt Optimization

Many works have developed iterative and black-box strategies that automatically refine prompts using task feedback. Some formulate the prompt optimization process as a search process, where LLMs iteratively refine instructions using prior solutions (Yang et al., 2024) or leverage Monte Carlo Tree Search (MCTS) to refine prompts from feedback (Wang et al., 2024b). Others apply policy-gradient methods to rewrite prompts for improved task accuracy (Kong et al., 2024) or propose an instruction-evolution framework that mutates and selects instructions over successive generations (Zeng et al., 2024). Similarly, RLPrompt (Deng et al., 2022) views prompt tokens as actions and learns a policy to generate high-reward prompts, while MAPO (Chen et al., 2023b) optimizes prompts via reinforcement learning, surpassing human-crafted prompts. ZO-Pog (Zhang et al., 2025) jointly tunes discrete and continuous prompts via policy gradient and zeroth-order optimization. Pryzant et al. (2023) formulate prompt optimization as a discrete analogue of gradient descent, combining self-critique with beam search. Zhou et al. (2022) extends this perspective with meta-prompting, where LLMs generate and rank instructions based on an evaluation signal, while Shin and Kim (2025) incorporates reasoning into prompt optimization to enhance prompt quality. Some works frame prompt optimization as an optimization problem analogous to gradient descent, offering a theoretical justification for the convergence of such

```
Your task is to write ONLY the {module_name} section of the
prompt for  stance detection towards the target {target}. Below
are the previous prompts with their scores. The score ranges
from 0 to 100, and higher scores indicate better quality
prompt.

{prompt_1}
{score_1}
...
{prompt_n}
{score_n}

Carefully analyze the previous prompts and their scores, and
write a new improved prompt to replace the prompt below:

{to_rewrite_prompt}

Wrap the new improved prompt between START and END tags. Do NOT
explain your changes or include any other text.
```

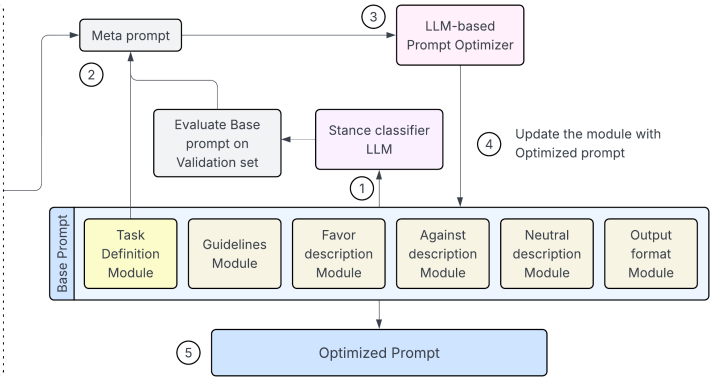


Figure 5.1: **Overview of the MoPrO-SD framework.** The process starts with a manually crafted base prompt, which is evaluated on a validation set. Each module is iteratively selected from the base prompt to be optimized using the meta-prompt. After optimization, the base prompt is updated with this new optimized module. The prompt optimizer then uses the next module to optimize, and so on. This iteration continues until the desired iteration, and the iteration with the best validation score is selected as the optimized prompt.

iterative methods (Tang et al., 2025; Trivedi et al., 2025).

Although prompt optimization has been widely studied in the context of LLMs, its application to stance detection remains limited. To bridge this gap, drawing on insights from prior studies on prompt optimization, we introduce a modular and iterative framework for prompt optimization, specifically tailored to the task of stance detection.

5.3 Methodology

In this section, we first discuss the our base prompt for stance detection, which consists of six distinct modules. Next, we introduce **Mo**dular **Pr**ompt **O**ptimization for **S**tance **D**etection (**MoPrO-SD**) framework, and finally, we discuss several optimization strategies, including *instruction-only*, *instruction with reasoning* and *instruction with per-example reasoning*. The overall methodology of our proposed framework is illustrated in Figure 5.1.

5.3.1 Stance detection prompt modules

We employ a modular prompt design for stance detection, starting with a human-written prompt that consists of six distinct modules. Treating an entire prompt as a single block often leads to fragmented and less effective refinements (Kumar et al., 2025), whereas a modular structure improves both optimization stability and overall performance (Khot et al., 2023; Tang et al., 2025). The *task definition* module states the objective of the task - specifying the task and goal, while the *guidelines* module specifies the decision criteria and highlights challenges posed by phenomena such as irony, sarcasm, and implicit stance. Three label-specific modules - *favor description*, *against description*, and *neutral description* - define descriptions for each class (optionally with representative examples) to reduce the boundary confusions that are common in stance detection tasks (Mohammad et al., 2016). Finally, an *output format* module enforces a consistent label-only response. This decomposition aligns with findings that LLMs benefit from explicit, structured instructions and reasoning decompositions in stance classification (Aiyappa et al., 2024; Ma et al., 2024) and leverages LLMs' capacity to iteratively optimize prompts when guided through targeted, module-level revisions (Tang et al., 2025; Zhou et al., 2022).

5.3.2 Proposed MoPrO-SD framework

MoPrO-SD employs an iterative, module-wise refinement strategy that treats the LLM itself as a prompt optimizer. The initial prompt for stance detection is created through manual prompt engineering and consists of the six modules, as described in the previous section. The complete prompt instructs an LLM to classify vaccine-related tweets into one of the three categories: *In-Favor*, *Against*, *Neutral* (see Appendix D.2 for an example of a complete prompt).

The initial six-module base prompt is evaluated on a validation set using a stance classifier LLM to obtain a baseline score, which represents the accuracy on the validation set. The optimization process then proceeds by updating one module at a time: a meta-prompt presents the LLM with the current version of a module, along with prior versions of the

prompt and their corresponding validation accuracy scores, and instructs it to propose an improved revision. The improved revision replaces the selected module while keeping the others fixed. After all modules are optimized, the updated prompt is re-evaluated on the validation set. This process is similar to retrieval-based trajectory methods in prompt optimization (Tang et al., 2025). Sequentially, cycling through all six modules constitutes one iteration, and the prompt version achieving the highest validation accuracy across iterations is selected as the final optimized prompt. The prompt optimizer could be the same LLM as the classifier or a more advanced LLM. As shown by Zhou et al. (2022), LLMs can autonomously generate and refine their own prompts, effectively serving as self-optimizers aligned with their characteristics and sensitivities.

Conceptually, this approach can be viewed as a coordinate descent in prompt space, where refinement is guided by the LLM's "reflection" on prior outcomes, consistent with findings in recent work on history-aware prompt optimization such as GPO (Tang et al., 2025). By independently optimizing each module, MoPrO-SD enables fine-grained prompt refinement that ultimately enhances the performance of LLM-based stance detection.

5.3.3 Methods of prompt optimization

The base prompt is composed of six modules, and each of these modules is individually optimized by the LLMs. We experiment with three types of iterative prompt optimization strategies for optimizing each module.

Instruction-only

In an instruction-only setting, we use a meta-prompt to direct the LLM to revise the module exclusively based on previous prompts and prompt scores. Thus, optimization is limited to lexical and syntactical reformulation of the text in the module. The optimizer LLM paraphrases the module to maximize the performance since LLMs are sensitive to subtle variations in wording and query phrasing (Zhou et al., 2022). See Appendix D.1 for an example of an instruction-only meta-prompt.

Instruction with reasoning

In the instruction with reasoning setting, we add an additional reasoning step, where we first ask the LLM to generate an explanation of how the favor/neutral/against modules could be improved, based on a subset of misclassified examples from the training set as evidence. Specifically, only instances in which the gold label belongs to a given class but the model predicts otherwise are included. For example, in refining the favor module, only samples with a true label of favor but misclassified by the model are selected. The meta-prompt then incorporates this reasoning response to refine the module. Utilizing this extra explanation of the classifier's errors allows the optimizer to fill in the instructions or details that might be missing in the module, making the module clearer and more effective (Shin and Kim, 2025). Appendix D.1 shows the details of the meta-prompt for the reasoning step and update.

Instruction with per-example reasoning

In the per-example reasoning, we optimize only the *favor*, *against*, or *neutral* description modules based on the ground truth of each training instance. First, we use the classifier to make predictions on a small batch of training examples and identify the instances where the model incorrectly classified the stance label. Following a similar method to Tang et al. (2025), whenever the current prompt gives a prediction that differs from the ground truth, a meta-prompt asks the LLM to explain why the instance was misclassified and provide an instruction on adjusting the corresponding module to classify the example correctly. These explanations capture subtle clues initially missed by the original descriptions. The optimizer then incorporates these insights and updates the module description by rewording, reordering, or adding relevant details. See Appendix D.1 for an example of the meta-prompt using instruction with per-example reasoning.

In the instruction-only setting, the same LLM serves as both the optimizer and the classifier. For tasks requiring reasoning, we evaluate two configurations: one using this same LLM as the reasoner, and another using GPT-5, a more advanced reasoning model from OpenAI (OpenAI, 2025).

Table 5.1: Examples of posts labeled by Llama3 model (llama_label) and human annotators (human_label)

Posts	llama_label	human_label
The autism one really gets me ... considering it's a genetic defect that you're born with and not something transmitted through vaccines.	in-favor	in-favor
Parents: heard the new vaccine requirement? Meningitis vaccine required for IL students entering 6th, 12th grade.	neither	neither
Then it's not recommended. the govt is making the public suffer by gambling with this, i cant. Make the non-medical govt officials take the vaccine first.	against	against
"One of the last holy grails of HIV research is the development of a HIV vaccine." #homenews #news #cryptonews	neither	in-favor
Watching the age go up for what's considered eligible for the vaccine. KNOCK IT OFF	against	in-favor

5.4 Experimental setup

5.4.1 Stance detection datasets

We used several existing stance detection benchmark datasets (SemEval-2016, GunStance, COVID-19-Stance), as well as a new VaccineStance dataset curated as part of this work.

1. **SemEval-2016 Task 6** The SemEval-2016 Task 6 dataset (Mohammad et al., 2016) is a benchmark dataset for stance detection, which consists of tweets that are manually annotated to indicate the author's stance towards a specific target. Each tweet is labeled as *Favor*, *Against*, or *None* (neutral/no expressed stance). The dataset consists of a total of 2,914 train instances and 1,249 test instances for 5 targets: *Atheism*, *Climate Change is a Real Concern*, *Feminist Movement*, *Hillary Clinton*, and *Legalization of Abortion*.

2. **GunStance** The GunStance dataset (Gyawali et al., 2024) is a benchmark dataset comprising manually annotated tweets related to gun control and gun regulation debates in the US. It focuses on two targets: banning guns and regulating guns. Each tweet is labeled as *Favor*, *Against*, or *Neutral*. The dataset consists a total of 1,836 train instances, 1,832 validation instances, and 1,832 test instances for two targets: *Banning guns* and *Regulating guns*. Additionally, this dataset contains unlabeled tweets

to facilitate semi-supervised learning.

3. **COVID-19-Stance** The COVID-19-Stance dataset (Glandt et al., 2021) is a benchmark dataset for stance detection in tweets related to the COVID-19 pandemic. The dataset comprises tweets that are manually annotated to indicate the author's stance towards specific pandemic-related targets. Each tweet is labeled as *Favor*, *Against*, or *None* (neutral/no expressed stance). The dataset contains 6,133 annotated tweets covering four targets: *Anthony S. Fauci, M.D.*, *Keeping Schools Closed*, *Stay at Home Orders*, and *Wearing a Face Mask*. Annotations were performed using Amazon Mechanical Turk, and only tweets with at least two annotators agreeing on the stance label were included. The dataset is split into 4,533 tweets for training, 800 for validation, and 800 for testing.

4. **VaccineStance** We collected historical public posts from X (formerly Twitter) using the platform's full archive search API, spanning a ten-year period from January 2013 to December 2022. This included seven years before COVID-19 (2013-2019) and three years during the pandemic (2020-2022). Using a comprehensive set of vaccine-related keywords, we gathered only English-language, original posts, excluding re-posts. After collection, we applied several preprocessing steps: retaining just one post per user per day, removing posts from prolific accounts, stripping re-posts, emoticons, URLs, non-ASCII characters, and user mentions, and filtering out posts related to the music band "The Vaccines" using a Sentence-Transformer (**?**) with a cosine similarity threshold of 0.7. This process resulted in 18.7 million posts from 5.8 million unique users.

   We sample 240,000 tweets (2000 tweets per month x 10 years) and use Llama3 (Meta) to classify if the tweet is relevant or irrelevant to the stance detection task towards vaccines. From the relevant tweets, we sample 2000 tweets and use human annotators to annotate them into three classes: *Favor*, *Against*, and *Neutral*. Each tweet is independently annotated by three annotators. For our entire set of 2000 tweets, the inter-coder agreement (Fleiss' kappa) was 0.778. We divide the dataset into 407 training, 578 validation, and 1015 test instances. Table 5.1 shows a few examples of tweets

from our dataset.

We remove *Donald Trump* target from the SemEval-2016 Task 6 dataset since it only had test instances and no training instances. Table A1 presents the overall class distribution across the datasets.

5.4.2 Large language models

For a comprehensive comparative analysis, we employ five open-source or open-weight large language models: Phi4-14b, Gpt-oss-20b, Mistral 3 Small, Gemma3-27B, and Llama 3.3. Descriptions of the Mistral 3 Small, Gemma3-27B, and Llama 3.3 models are presented in Chapter 4, Section 4.4.1. The descriptions of the remaining two Phi-4 and Gpt-oss-20B models are presented below:

1. **Microsoft Phi-4:14B:** The Phi-4 model (Phi4:14B) is an open-weight 14-billion-parameter, text-only, instruction-tuned model optimized for reasoning tasks (Abdin et al., 2024). It uses a dense decoder-only Transformer architecture with a 16k-token context window and is trained on a mixture of high-quality synthetic data and filtered sources (e.g., public-domain web documents, licensed books, and Q&A datasets). The models' post-training combines supervised fine-tuning and iterative Direct Preference Optimization to improve instruction adherence and safety. On standard evaluations, Phi-4 (14B) improves over Phi-3 (14B) model and matches or exceeds larger models (e.g., GPT-4o) on several reasoning benchmarks (e.g., GPQA and MATH), while remaining suitable for memory/latency-constrained deployments (Abdin et al., 2024).

2. **Gpt-oss:20b:** The gpt-oss-20b model is an open-weight, instruction-tuned, and text-only generative model from OpenAI (2025), designed for reasoning and tool use. It is based on the Mixture-of-Experts (MoE) Transformer architecture with 21B total parameters and 3.6B active parameters per token that contains 32 experts, and only 4 are active for each token. For efficient inference, it uses grouped multi-query attention and supports a long-context window of over 128,000 tokens. Post-training step combines

the supervised fine-tuning with reinforcement-learning-based alignment. On standard evaluations benchmarks, gpt-oss-20b achieves performance comparable to OpenAI's o3-mini, while remaining deployable on devices with around 16 GB memory (OpenAI, 2025).

5.4.3 Research questions

We run a comprehensive set of experiments to evaluate the proposed framework and understand what models and strategies may give the best results. Specifically, our experiments are designed to answer the following questions:

RQ1 How does the validation accuracy vary with the number of optimization iterations for different datasets and different models?

RQ2 Given the datasets and models considered, what setting leads to better results - zero-shot or few-shot?

RQ3 Overall, what model performs the best across the four datasets used in our study?

RQ4 Among the three strategies used for optimization, instruction-only (I), instruction with reasoning (I+R), and instruction with per-example reasoning (I+R+E), which strategy is the best across datasets?

RQ5 Given that reasoning can be done with the base model itself, or with a more powerful reasoning model such as GPT-5, which of these two options gives better results?

5.5 Results and discussion

Figure 5.2 shows the accuracy of various models during optimization steps using the MoPrO-SD framework and can be used to answer RQ1. As can be seen, in general, accuracy tends to improve as the optimization proceeds. While occasional drops occur at certain steps, later stages usually recover and exceed the baseline performance. The optimal prompt is identified

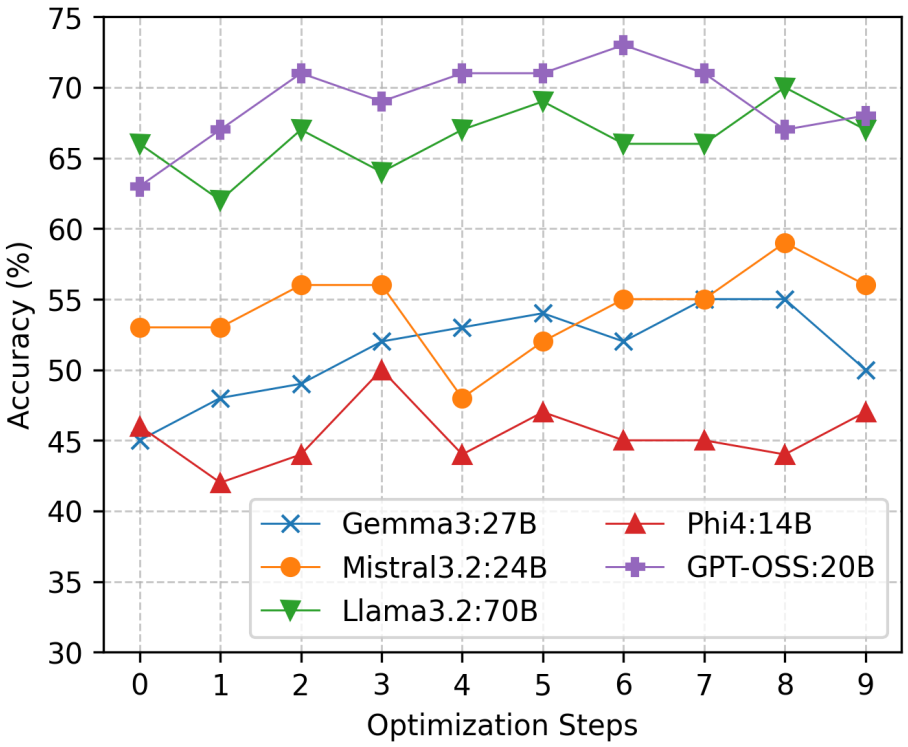


Figure 5.2: Validation accuracy of different models over 10 optimization iterations on the VaccineStance dataset using instruction-only, zero-shot optimization.

as the one that achieves the highest validation score. Among the models, GPT-OSS:20B and Gemma3:27B display steady and consistent improvement, whereas Mistral3.2:24B and Phi4:14B exhibit less stable progress across the optimization steps.

To answer RQ2, in Figure 5.3 we show the accuracy of different models using optimized prompts in both zero-shot and few-shot settings across the four datasets. For the VaccineStance dataset, few-shot prompting consistently outperforms zero-shot prompting. However, this pattern does not hold uniformly across the other datasets. In the GunStance dataset, both Mistral3.2:24B and Phi4:14B achieve higher accuracy with few-shot prompting, while in the SemEval-2016 dataset, only the Phi4:14B model shows improvement under the few-shot setting. These results suggest that the advantage of few-shot prompting is dataset- and model-dependent, rather than universally consistent.

To answer research questions RQ3 and RQ4, for each dataset and each model, we evaluate the MoPrO-SD framework at the target level using three prompt settings: Instruction-only

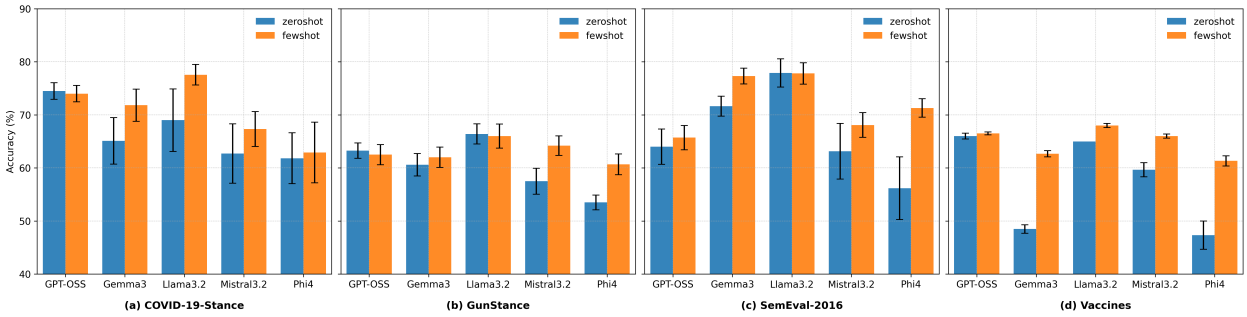


Figure 5.3: Performance evaluation of zero-shot and few-shot prompting across four benchmark datasets: (a) COVID-19-Stance, (b) GunStance, (c) SemEval-2016, and (d) VaccineStance.

(I), Instruction with reasoning (I+R), and Instruction with per-example reasoning (I+R+E).

Table 5.2 shows the performance of various models across five targets present in the SemEval-2016 dataset. We find the performance varies by targets. Llama3.3:70B is generally the best-performing model, showing 77.40 ± 1.82 with I+R on *atheism* target and 86.0 ± 1.0 with I+R+E on *Hillary Clinton* target. It also shows noticeable gain on *climate change is a real concern* target with I+R+E (85.50 ± 0.71 as compared to 83.16 ± 1.92 with base prompt). In contrast, for *feminist movement* target, the base prompt scores the best across all models, indicating that added reasoning or optimization can be unnecessary or even harmful. For the *legalization of abortion* target, Gemma3:27B is the best performing model with I+R+E (72.33 ± 1.53).

Similarly, Table 5.3 shows the performance of various models across the two targets present in the GunStance dataset. Across both targets–*guns should be banned* and *guns should be regulated*–Llama3.2:70B consistently scores the best score as compared to all models. It performs best under I or I+R setting (e.g., 62.25 ± 1.89 on *guns should be banned* target with I+R; and 71.20 ± 1.64 on *guns should be regulated* target with I+R). I+R yields modest but consistent gains over the base prompt for most models, while I+R+E degrades the performance (notably for Mistral3.2:24B and Phi4:14B models). Overall, instruction-only optimization and instruction with reasoning are helpful, while per-example reasoning does not provide much improvement for this dataset.

Table 5.4 shows the performance of various models across the five targets present in the

Table 5.2: Results on the SemEval-2016 dataset. I = instruction only; I+R = instruction + reasoning; I+R+E = instruction + reasoning + per-example reasoning. Base = original prompt. Green cells indicate an improvement over the Base score, while red cells indicate a decrease. Darker shades represent a larger absolute difference from the Base. The best score for the target is **highlighted**.

		Model				
Target	**Method**	**Phi4:14B**	**Gpt-oss:20B**	**Mistral3.2:24B**	**Gemma3:27B**	**Llama3.2:70B**
Atheism	Base	43.54 ± 15.01	56.66 ± 10.37	31.03 ± 11.32	70.90 ± 2.65	72.45 ± 4.20
	I	50.67 ± 14.98	62.00 ± 9.20	53.00 ± 9.54	71.40 ± 4.10	73.33 ± 4.93
	I+R	47.50 ± 17.68	69.67 ± 3.33	63.33 ± 11.55	72.67 ± 2.50	**77.40 ± 1.82**
	I+R+E	33.00 ± 1.00	70.00 ± 3.94	32.00 ± 18.03	73.67 ± 5.51	73.33 ± 8.33
Climate change is a real concern	Base	42.78 ± 1.16	68.66 ± 0.94	60.50 ± 0.93	64.86 ± 4.62	83.16 ± 1.92
	I	47.25 ± 5.91	67.00 ± 3.52	58.25 ± 9.18	72.83 ± 5.64	82.80 ± 2.59
	I+R	35.00 ± 7.07	68.40 ± 7.4	81.00 ± 6.24	82.17 ± 3.19	84.00 ± 1.00
	I+R+E	39.00 ± 2.00	72.00 ± 6.22	61.33 ± 7.64	68.50 ± 10.61	**85.50 ± 0.71**
Feminist movement	Base	64.36 ± 4.21	66.55 ± 0.39	59.21 ± 4.36	75.03 ± 0.48	**77.00 ± 1.97**
	I	66.20 ± 4.66	66.50 ± 0.71	67.50 ± 4.04	75.00 ± 1.00	75.40 ± 2.07
	I+R	63.67 ± 5.13	68.00 ± 1.67	67.75 ± 5.74	70.20 ± 5.31	75.00 ± 0.82
	I+R+E	52.67 ± 10.02	65.00 ± 1.14	58.67 ± 14.57	73.00 ± 2.83	75.00 ± 1.00
Hillary Clinton	Base	73.67 ± 0.00	76.00 ± 0.95	73.67 ± 1.61	81.72 ± 0.35	83.36 ± 1.20
	I	73.33 ± 1.53	76.00 ± 1.67	75.75 ± 2.75	82.00 ± 0.82	84.50 ± 1.29
	I+R	76.33 ± 4.51	74.50 ± 2.88	76.67 ± 1.53	80.60 ± 2.07	83.00 ± 2.35
	I+R+E	74.00 ± 4.21	74.00 ± 1.26	61.67 ± 13.65	83.00 ± 2.65	**86.00 ± 1.00**
Legalization of abortion	Base	59.00 ± 2.27	60.16 ± 1.18	63.37 ± 1.13	70.55 ± 0.93	69.82 ± 3.03
	I	57.67 ± 5.51	56.00 ± 1.82	64.00 ± 1.83	71.33 ± 1.53	69.33 ± 2.08
	I+R	58.00 ± 1.41	58.40 ± 1.82	65.33 ± 3.79	70.60 ± 1.14	69.20 ± 0.84
	I+R+E	59.50 ± 0.71	61.00 ± 2.30	64.00 ± 7.21	**72.33 ± 1.53**	70.00 ± 1.00

Table 5.3: Results on the GunStance dataset. I = instruction only; I+R = instruction + reasoning; I+R+E = instruction + reasoning + per-example reasoning. Base = original prompt. Green cells indicate an improvement over the Base score, while red cells indicate a decrease. Darker shades represent a larger absolute difference from the Base. The best score for the target is **highlighted**.

Target	**Method**	**Phi4:14B**	**Gpt-oss:20B**	**Mistral3.2:24B**	**Gemma3:27B**	**Llama3.2:70B**
Guns should be banned	Base	50.79 ± 2.93	57.27 ± 0.37	49.44 ± 7.67	50.27 ± 5.61	58.45 ± 1.16
	I	54.00 ± 5.57	59.00 ± 2.36	54.50 ± 6.14	58.40 ± 2.70	61.75 ± 1.26
	I+R	50.00 ± 5.66	57.67 ± 2.31	57.00 ± 7.07	61.00 ± 2.65	**62.25 ± 1.89**
	I+R+E	44.33 ± 1.53	56.00 ± 1.73	47.00 ± 8.19	58.00 ± 2.00	60.00 ± 0.82
Guns should be regulated	Base	51.96 ± 3.90	65.42 ± 0.83	52.61 ± 4.76	60.00 ± 4.83	69.72 ± 1.12
	I	54.00 ± 7.07	67.00 ± 0.58	60.00 ± 1.41	67.33 ± 2.08	71.00 ± 1.58
	I+R	54.00 ± 8.49	66.50 ± 2.12	64.00 ± 5.66	68.20 ± 1.30	**71.20 ± 1.64**
	I+R+E	45.67 ± 1.53	67.00 ± 1.21	54.50 ± 3.54	67.00 ± 1.41	69.67 ± 2.08

COVID-19-Stance dataset. On the *face masks* target, Llama3.2:70B attains the best score overall with instruction with reasoning setting (87 ± 2.65), improving markedly over the base accuracy of 81.97 ± 1.35. Similarly, *school closure* shows the largest relative jump for

Table 5.4: Results on the COVID-19-Stance dataset. I = instruction only; I+R = instruction + reasoning; I+R+E = instruction + reasoning + per-example reasoning. Base = original prompt. Green cells indicate an improvement over the Base score, while red cells indicate a decrease. Darker shades represent a larger absolute difference from the Base. The best score for the target is **highlighted**.

Target	Method	Phi4:14B	Gpt-oss:20B	Mistral3.2:24B	Gemma3:27B	Llama3.2:70B
Anthony Fauci	Base	66.33 ± 2.74	73.22 ± 0.19	70.60 ± 1.20	63.72 ± 5.63	70.76 ± 1.51
	I	72.00 ± 1.41	76.50 ± 3.54	71.00 ± 2.16	67.00 ± 5.76	72.00 ± 3.16
	I+R	71.00 ± 3.03	75.33 ± 2.8	73.75 ± 1.71	72.80 ± 3.03	72.40 ± 2.51
	I+R+E	74.00 ± 1.00	76.00 ± 1.64	73.00 ± 1.41	67.00 ± 7.00	**74.50 ± 2.12**
Face masks	Base	63.33 ± 3.61	76.17 ± 1.66	70.78 ± 3.17	72.00 ± 3.38	81.97 ± 1.35
	I	54.33 ± 17.62	75.00 ± 1.41	69.67 ± 3.06	74.67 ± 1.63	82.00 ± 1.58
	I+R	66.50 ± 2.12	77.00 ± 1.26	77.50 ± 0.58	79.00 ± 3.35	**87.00 ± 2.65**
	I+R+E	69.00 ± 3.21	77.00 ± 0.58	71.00 ± 2.83	78.33 ± 3.06	86.33 ± 4.73
School closures	Base	44.78 ± 5.01	71.78 ± 1.35	47.82 ± 9.66	45.47 ± 11.16	53.43 ± 16.37
	I	48.00 ± 3.94	72.00 ± 1.41	47.75 ± 9.74	55.50 ± 6.09	55.75 ± 17.69
	I+R	67.00 ± 5.57	69.50 ± 2.08	60.00 ± 7.81	**72.17 ± 6.97**	72.00 ± 13.39
	I+R+E	48.00 ± 2.12	71.00 ± 3.27	52.00 ± 4.24	71.00 ± 7.21	61.33 ± 9.71
Stay at home orders	Base	74.84 ± 2.12	80.67 ± 0.00	74.78 ± 2.17	77.45 ± 2.62	78.31 ± 1.82
	I	78.00 ± 0.58	80.50 ± 0.71	75.00 ± 2.65	80.75 ± 2.99	77.83 ± 2.64
	I+R	72.50 ± 0.71	78.60 ± 4.39	79.75 ± 0.96	**82.00 ± 2.45**	80.20 ± 1.92
	I+R+E	80.00 ± 2.17	78.00 ± 2.17	77.00 ± 4.24	80.67 ± 1.53	79.00 ± 1.00

Gemma3:27B model, improving the performance from 45.47 ± 11.16 to 72.17 ± 6.97, while GPT-OSS:20B remains strong without added reasoning (Base 71.78 ± 1.35; I 72.0 ± 1.41). For *stay at home orders* target, Gemma3:27B scores the highest with I+R (82.0 ± 2.45). Phi4:14B improves only modestly across all targets.

Finally, Table 5.5 shows the performance of various models across our dataset for stance detection on *vaccines*. Llama3.2:70B performs the best using the base prompt among all the models with the same base prompt, whereas Gemma3:27B model scores the highest with 69.20 ± 1.92 with I+R optimization among all models and optimization strategy. I+R+E does not help much and often hurts the performance across models (e.g., Gemma3:27B scores 54.67 ± 7.02). GPT-OSS:20B performs well with instruction-only optimization, while Phi4:14B remains comparatively weak across all settings for this dataset.

Optimizing instructions only (I) provides consistent improvement over the baseline, while adding optimizing instructions with reasoning frequently yields the best scores (especially for Llama3.2:70B and Gemma3:27B models). Per-example reasoning-based instruction optimization is rarely beneficial and can reduce stability and accuracy. Our results show that the

Table 5.5: Results on the Vaccines dataset. I = instruction only; I+R = instruction + reasoning; I+R+E = instruction + reasoning + per-example reasoning. Base = original prompt. Green cells indicate an improvement over the Base score, while red cells indicate a decrease. Darker shades represent a larger absolute difference from the Base. The best score for the target is **highlighted**.

Target	Method	Phi4:14B	Gpt-oss:20B	Mistral3.2:24B	Gemma3:27B	Llama3.2:70B
Vaccines	Base	50.78 ± 7.57	62.67 ± 1.56	54.73 ± 6.56	49.08 ± 10.63	66.00 ± 1.04
	I	51.75 ± 9.78	65.00 ± 0.96	62.75 ± 4.79	57.00 ± 7.91	67.25 ± 1.71
	I+R	54.00 ± 10.65	66.50 ± 1.29	61.00 ± 1.83	68.20 ± 1.92	**69.80 ± 0.84**
	I+R+E	50.00 ± 2.65	65.40 ± 2.07	58.33 ± 1.53	54.67 ± 7.02	65.80 ± 2.59

effects of optimization strategies vary considerably across targets and models. However, the MoPro-SD framework with instruction-only optimization consistently demonstrates robust and consistent advantages relative to the base prompts.

Related to RQ5, Table 5.6 shows the overall scores of all models on the benchmark datasets. In these experiments, GPT-5 (OpenAI, 2025), a flagship model from OpenAI, is employed as the reasoning model. For comparison, we also evaluate the case where the optimizer and the reasoning model are the same. Across all datasets, the results show that using the larger and more capable GPT-5 as the reasoning model leads to substantially higher performance than relying on the same model in both roles.

5.6 Conclusion

In this work, we introduced MoPrO-SD, a modular prompt optimization framework designed for stance detection. By decomposing a long prompt into functional modules and optimizing them iteratively, the framework provides an automated prompt optimization. Empirical results across four datasets–including our newly curated vaccines dataset–demonstrate consistent improvements over the base prompts. Among the optimization strategies explored, refining the modules based on the instructions alone, or in combination with reasoning, yielded the most stable and substantial gains, whereas optimization based on per-example reasoning was less reliable and often detrimental. Moreover, utilizing a stronger external model as a reasoner further enhanced the effectiveness of optimization. These findings sug-

Table 5.6: Accuracy comparison of prompt optimization across datasets using two reasoning models. "Same" denotes using the same model as the stance classifier for reasoning, while "GPT-5" denotes using the GPT-5 model for reasoning. I+R indicates prompts optimized with instructions plus reasoning, and I+R+E indicates prompts optimized with per-example reasoning. The baseline refers to the accuracy of the original prompt. Green cells indicate an improvement over the Base score, while red cells indicate a decrease. Darker shades represent a larger absolute difference from the Base. The best score for the target is **highlighted**.

Dataset	Baseline	Same		GPT-5	
		I+R	I+R+E	I+R	I+R+E
COVID-19-Stance	65.03	75.15	74.20	**76.37**	75.82
GunStance	54.60	62.35	61.88	**64.57**	63.70
SemEval-2016	64.52	72.22	70.52	**72.70**	71.92
Vaccines	58.90	64.47	63.40	**69.40**	67.53

gest that making edits to long, complex prompts by decomposing them into smaller modules, guided by small validation sets, can push the performance of large language models without fine-tuning their parameters. Importantly, MoPrO-SD offers a practical and effective means of enhancing the prompt performance, particularly for closed or API-based models.

Limitations

Our work has a few limitations that must be acknowledged. First, the generalizability of our framework beyond the stance detection domain remains unexplored. In addition, the optimal prompt tends to be model-specific, and a new LLM or domain might require re-optimization. Furthermore, the human annotation strategy for stance may introduce subjective biases, and the impact of these biases on prompt optimization is not fully examined; addressing this remains part of our future work. Finally, our framework is tailored specifically to stance detection with fixed favor/against/neutral modules, and extending its applicability to other tasks or more complex label schemas will be pursued in future research.

Chapter 6

Concluding remarks and future prospects

The task of Stance detection has broad applications in various domains, from misinformation detection and public opinion mining to extracting key financial performance metrics from financial documents. The proliferation of digital platforms has yielded vast, heterogeneous textual data. This abundance and variety of generated information offer rich opportunities for applying stance detection methodologies to uncover valuable insights for understanding and addressing critical challenges in today's modern world. The emergence of Large Language Models (LLMs) has transformed the field of NLP by offering unprecedented capabilities in understanding subtle linguistic cues and context-dependent meanings and addressing the limitations of traditional stance detection methodologies.

We set out to answer a simple question with far-reaching consequences–given a text and a target or claim, can we classify if the text is in favor, against, or neutral towards the target or claim? We saw that this task, although seemingly easy, has a lot of challenges. Often texts carry implicit meanings along with sarcasm, irony, numerically dense prose, and polite wordings that can hide opinions, making stance detection challenging for models. This dissertation approached stance detection as a structured reasoning problem and developed techniques that utilize the power of Large Language Models (LLMs) to make the task more

reliable, scalable, and economical in terms of dataset annotation.

In Chapter 2, we introduced GunStance, a stance detection dataset collected around major shooting events in the US, and proposed a hybrid semi-supervised method with LLMs that leverages the unlabeled data to improve the performance while reducing the data annotation costs.

We also constructed a decade-long dataset of vaccine discourse on X (formerly Twitter), spanning over pre-COVID-19 (2013-2019) and COVID-19 (2020-2022) years in Chapter 3. Using lexicon-based emotion analysis and stance classification, we showed how language regarding vaccines moved not only in positive/negative dimensions but also along a "warmth/-competence" dimension, offering a more nuanced discourse on how people talk about vaccines over time.

Similarly, in Chapter 4, we created a sentence-level corpus for financial stance detection towards specific metrics–debt, EPS, and sales–from SEC 10-K reports and earnings call transcripts (ECTs). With LLM-assisted labeling with careful human validation, we systematically compared prompting strategies for various LLMs. We demonstrated that few-shot prompting with chain-of-thought reasoning tends to outperform zero-shot prompting, and performance depends on the source (formal SEC prose vs. conversational ECT).

Finally, we proposed Modular Prompt Optimization for Stance Detection (MoPrO-SD), a framework that decomposes the long stance detection prompt into smaller modules and improves each module iteratively by leveraging LLMs with meta-prompts in Chapter 5. This framework delivered consistent performance gains across models and datasets.

Across domains and methodologies, we learned that:

1. Stance detection is target-dependent and context-sensitive. Separating closely related targets (e.g., banning guns vs regulating guns) and sentences that across financial metrics (e.g., debt, EPS, sales) makes the models robust without massive labels.

2. LLMs can perform well under careful prompting strategies like Few-shot demonstrations and chain-of-thought reasoning.

3. Decomposing a long, complex prompt into modular components and employing LLM to

iteratively optimize each module frequently produces prompts that outperform purely human-crafted prompt baselines.

There are a few limitations that must be kept in mind when developing stance detection methodologies. Social media corpora reflect the users of the platform, not the broader population. What is visible on social media often depends on posting incentives, the presence of prolific users, and evolving API access. Similarly, stance labels–either human or LLM-assisted–encode contemporary norms, and they may drift over time. LLM-annotated data may encode biases from the LLMs. Financial texts vary by industry and time period. Models tuned on one company may not transfer perfectly to another. While we reduce some limitations through human validation of LLM-assisted labels and the use of inter-annotator agreement to quantify uncertainty, the core limitations of social data drift and differences in data across domains still persist.

Building upon our results and findings, several paths look promising for future work. Firstly, extending from a text-based stance detection to multimodal stance detection by leveraging images, tables, and charts could aid models to perform better is a natural next step. Many social media posts pair texts with photos, memes, or videos, and financial disclosures often embed key insights in tables and figures. Similarly, we can apply the MoPrO-SD framework beyond the stance detection task, such as claim verification, misinformation detection, and financial risk and outlook classification by swapping the modules based on the tasks.

In summary, the task of stance detection will never be "done" as language and norms keep evolving. In this dissertation, we showed that with the right ingredients–curation of target-specific datasets, crafting careful prompts with few-shot examples with chain-of-thought reasoning demonstrations, and modular prompt optimization–LLMs can deliver reliable, scalable, and cost-effective stance classification. Together, these contributions offer a framework for informed and nuanced decision-making in an increasingly digital and interconnected world.

Bibliography

M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

A. E. Abele and B. Wojciszke. Communal and agentic content in social cognition: A dual perspective model. In *Advances in experimental social psychology*, volume 50, pages 195–255. Elsevier, 2014.

A. E. Abele, N. Hauke, K. Peters, E. Louvet, A. Szymkow, and Y. Duan. Facets of the fundamental content dimensions: Agency with competence and assertiveness—communion with warmth and morality. *Frontiers in psychology*, 7:1810, 2016.

J. Aggarwal, E. Rabinovich, and S. Stevenson. Exploration of gender differences in COVID-19 discourse on Reddit. In K. Verspoor, K. B. Cohen, M. Dredze, E. Ferrara, J. May, R. Munro, C. Paris, and B. Wallace, editors, *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.nlpcovid19-acl.13/.

R. Aiyappa, S. Senthilmani, J. An, H. Kwak, and Y.-Y. Ahn. Benchmarking zero-shot stance detection with flant5-xxl: Insights from training data, prompting, and decoding strategies into its near-sota performance. *arXiv preprint arXiv:2403.00236*, 2024.

S. Alahmadi, R. Hoyle, M. Head, and M. Brede. Modelling the mitigation of anti-vaccine opinion propagation to suppress epidemic spread: A computational approach. *PloS one*, 20(3):e0318544, 2025.

A. ALDayel and W. Magdy. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597, 2021.

H. Alhuzali, T. Zhang, and S. Ananiadou. Emotions and topics expressed on twitter during the covid-19 pandemic in the united kingdom: Comparative geolocation and text mining analysis. *Journal of Medical Internet Research*, 24(10), 2022. URL https://www.jmir.org/2022/10/e40323.

E. Allaway and K. McKeown. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931, Online, Nov. 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.emnlp-main.717.

N. S. Alturayeif, H. A. Luqman, and M. A. K. Ahmed. Mawqif: A multi-label arabic dataset for target-specific stance detection. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 174–184, 2022.

A. Amel-Zadeh and J. Faasse. The information content of 10-k narratives: comparing md&a and footnotes disclosures. *Available at SSRN 2807546*, 2016.

W. Antweiler and M. Z. Frank. Is all that talk just noise? the information content of internet stock message boards. *The Journal of finance*, 59(3):1259–1294, 2004.

D. Araci. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.

O. Araque, L. Gatti, and K. Kalimeri. Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-based systems*, 191: 105184, 2020.

O. Araque, L. Gatti, and K. Kalimeri. Libertymfd: A lexicon to assess the moral foundation of liberty. In *Proceedings of the 2022 ACM Conference on Information Technology for Social Good*, pages 154–160, 2022.

I. Augenstein, T. Rocktäschel, A. Vlachos, and K. Bontcheva. Stance detection with bidirectional conditional encoding. In J. Su, K. Duh, and X. Carreras, editors, *Proceedings*

of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 876–885, Austin, Texas, Nov. 2016. Association for Computational Linguistics. URL https://aclanthology.org/D16-1084/.

A. Bandura. *Social cognitive theory: An agentic perspective on human nature*. John Wiley & Sons, 2023.

G. V. Bodenhausen, S. K. Kang, and D. Peery. Social categorization and the perception of social groups. *The Sage handbook of social cognition*, pages 311–329, 2012.

Z. Bozanic, J. R. Dietrich, and B. A. Johnson. Sec comment letters and firm disclosure. *Journal of Accounting and Public Policy*, 36(5):337–357, 2017.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

E. Burwell, A. Agarwal, and W. L. Romine. Understanding communication about the COVID-19 vaccines: analysis of emergent sentiments and topics of discussion on Twitter during the initial phase of the vaccine rollout. *International Journal of Science Education, Part B*, 14(1):18–46, Jan. 2024. ISSN 2154-8455, 2154-8463. URL https://www.tandfonline.com/doi/full/10.1080/21548455.2023.2185829.

B. J. Bushee, D. A. Matsumoto, and G. S. Miller. Open versus closed conference calls: the determinants and effects of broadening access to disclosure. *Journal of accounting and economics*, 34(1-3):149–180, 2003.

R. A. Cazier and R. J. Pfeiffer. Why are 10-k filings so long? *Accounting Horizons*, 30(1): 1–21, 2016.

J. Chen, L. Chen, H. Huang, and T. Zhou. When do you need chain-of-thought prompting for chatgpt? *arXiv preprint arXiv:2304.03262*, 2023a.

Y. Chen, Z. Wen, G. Fan, Z. Chen, W. Wu, D. Liu, Z. Li, B. Liu, and Y. Xiao. Mapo: Boosting large language model performance with model-adaptive prompt optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3279–3304, 2023b.

R. Chiong, Z. Fan, Z. Hu, M. T. Adam, B. Lutz, and D. Neumann. A sentiment analysis-based machine learning approach for financial market prediction via news disclosures. In *Proceedings of the genetic and evolutionary computation conference companion*, pages 278–279, 2018.

E. Cicekyurt and G. Bakal. Enhancing sentiment analysis in stock market tweets through bert-based knowledge transfer. *Computational Economics*, pages 1–23, 2025.

M. Cinelli, G. De Francisci Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118, 2021.

C. Conforti, J. Berndt, M. T. Pilehvar, C. Giannitsarou, F. Toxvaerd, and N. Collier. Will-they-won't-they: A very large dataset for stance detection on Twitter. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1724, Online, July 2020a. Association for Computational Linguistics. URL https://aclanthology.org/2020.acl-main.157.

C. Conforti, J. Berndt, M. T. Pilehvar, C. Giannitsarou, F. Toxvaerd, and N. Collier. Will-they-won't-they: A very large dataset for stance detection on Twitter. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1724, Online, July 2020b. Association for Computational Linguistics. URL https://aclanthology.org/2020.acl-main.157/.

A. J. Cuddy, P. Glick, and A. Beninger. The dynamics of warmth and competence judgments, and their outcomes in organizations. *Research in organizational behavior*, 31:73–98, 2011.

E. D'Andrea, P. Ducange, A. Bechini, A. Renda, and F. Marcelloni. Monitoring the public opinion about the vaccination topic from tweets analysis. *Expert Systems with Applications*, 116:209–226, Feb. 2019. ISSN 09574174. URL https://linkinghub.elsevier.com/retrieve/pii/S0957417418305803.

S. De Rosis, M. Lopreite, M. Puliga, and M. Vainieri. The early weeks of the Italian Covid-19 outbreak: sentiment insights from a Twitter analysis. *Health Policy*, 125(8):987–994, Aug. 2021. ISSN 01688510. URL https://linkinghub.elsevier.com/retrieve/pii/S0168851021001627.

M. Deng, J. Wang, C.-P. Hsieh, Y. Wang, H. Guo, T. Shu, M. Song, E. Xing, and Z. Hu. Rlprompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3369–3391, 2022.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

J. Du, R. Xu, Y. He, and L. Gui. Stance classification with target-specific neural attention networks. In *26th International Joint Conference on Artificial Intelligence, IJCAI 2017*, IJCAI'17, page 3988–3994. AAAI Press, 2017. ISBN 9780999241103.

K. Du, F. Xing, R. Mao, and E. Cambria. Financial sentiment analysis: Techniques and applications. *ACM Computing Surveys*, 56(9):1–42, 2024.

K. Du, Y. Zhao, R. Mao, F. Xing, and E. Cambria. Natural language processing in finance: A survey. *Information Fusion*, 115:102755, 2025.

A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

F. Durante and S. T. Fiske. How social-class stereotypes maintain inequality. *Current opinion in psychology*, 18:43–48, 2017.

S. Ebner, F. Wang, and B. Van Durme. Bag-of-words transfer: Non-contextual techniques for multi-task learning. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 40–46, 2019.

P. Ekman. Facial expressions. *The Handbook of Cognition and Emotion/John Wiley & Sons*, 16(301):e320, 1999.

G. Fatouros, J. Soldatos, K. Kouroumali, G. Makridis, and D. Kyriazis. Transforming sentiment analysis in the financial domain with chatgpt. *Machine Learning with Applications*, 14:100508, 2023. ISSN 2666-8270. URL https://www.sciencedirect.com/science/article/pii/S2666827023000610.

Z. Feng, G. Hu, B. Li, and J. Wang. Unleashing the power of ChatGPT in finance research: opportunities and challenges. *Financial Innovation*, 11(1):93, Mar. 2025. ISSN 2199-4730. doi: 10.1186/s40854-025-00770-3. URL https://doi.org/10.1186/s40854-025-00770-3.

S. Fiske, A. Cuddy, P. Glick, and J. Xu. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, 82:878–902, 06 2002. URL https://www.doi.org/10.1037/0022-3514.82.6.878.

S. T. Fiske. Stereotype content: Warmth and competence endure. *Current directions in psychological science*, 27(2):67–73, 2018.

S. T. Fiske and F. Durante. Stereotype content across cultures. *Handbook of advances in culture and psychology*, 6:209–258, 2016.

S. T. Fiske, A. J. Cuddy, and P. Glick. Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, 11(2):77–83, 2007.

S. T. Fiske, F. Durante, et al. Never trust a politician? collective distrust, relational accountability, and voter response. *Power, politics, and paranoia: Why people are suspicious of their leaders*, pages 91–105, 2014.

S. T. T. Fiske. *Social cognition: From brains to culture*. SAGE Publications Ltd, 2020.

S. Fournier and C. Alvarez. Brands as relationship partners: Warmth, competence, and in-between. *Journal of consumer psychology*, 22(2):177–185, 2012.

Y. Fu, X. Li, Y. Li, S. Wang, D. Li, J. Liao, and J. Zheng. Incorporate opinion-towards for stance detection. *Knowledge-Based Systems*, 246:108657, 2022.

M. Gambini, C. Senette, T. Fagni, and M. Tesconi. Evaluating large language models for user stance detection on x (twitter). *Machine Learning*, pages 1–24, 2024.

S. Ghosh, P. Singhania, S. Singh, K. Rudra, and S. Ghosh. Stance detection in web and social media: a comparative study. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10*, pages 75–87. Springer, 2019.

M. D. Giovanni, F. Pierri, C. Torres-Lugo, and M. Brambilla. VaccinEU: COVID-19 Vaccine Conversations on Twitter in French, German and Italian. *Proceedings of the International AAAI Conference on Web and Social Media*, 16:1236–1244, May 2022. ISSN 2334-0770, 2162-3449. URL https://ojs.aaai.org/index.php/ICWSM/article/view/19374.

K. Glandt, S. Khanal, Y. Li, D. Caragea, and C. Caragea. Stance detection in covid-19 tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Long Papers)*, volume 1, 2021.

M. Gole, W.-P. Nwadiugwu, and A. Miranskyy. On sarcasm detection with openai gpt-based models. In *2024 34th International Conference on Collaborative Advances in Software and COmputiNg (CASCON)*, pages 1–6. IEEE, 2024.

J. Graham, J. Haidt, S. Koleva, M. Motyl, R. Iyer, S. P. Wojcik, and P. H. Ditto. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier, 2013.

A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

I. Gül, R. Lebret, and K. Aberer. Stance detection on social media with fine-tuned large language models. *arXiv preprint arXiv:2404.12171*, 2024.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Y. Guo, Z. Xu, and Y. Yang. Is chatgpt a financial expert? evaluating language models on financial natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 815–821, 2023.

N. Gyawali, I. Sirbu, T. Sosea, S. Khanal, D. Caragea, T. Rebedea, and C. Caragea. Gunstance: Stance detection for gun control and gun regulation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12027–12044, 2024.

O. Hamad, A. Hamdi, S. Hamdi, and K. Shaban. Steducov: an explored and benchmarked dataset on stance detection in tweets towards online education during covid-19 pandemic. *Big Data and Cognitive Computing*, 6(3):88, 2022.

E. L. Hamaker and M. Wichers. No Time Like the Present: Discovering the Hidden Dynamics in Intensive Longitudinal Data. *Current Directions in Psychological Science*, 26(1):10–15, 2017. URL https://doi.org/10.1177/0963721416666518.

T. A. Hassan, S. Hollander, A. Kalyani, L. van Lent, M. Schwedeler, and A. Tahoun. Economic surveillance using corporate text. Technical report, National Bureau of Economic Research, 2024.

Z. He, N. Mokhberian, and K. Lerman. Infusing knowledge from Wikipedia to enhance stance detection. In J. Barnes, O. De Clercq, V. Barriere, S. Tafreshi, S. Alqahtani, J. Sedoc, R. Klinger, and A. Balahur, editors, *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 71–77, Dublin, Ireland, May 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wassa-1.7/.

G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

W. E. Hipson and S. M. Mohammad. Emotion dynamics in movie dialogues. *PLOS ONE*, 16(9):e0256153, Sept. 2021. ISSN 1932-6203. URL https://dx.plos.org/10.1371/journal.pone.0256153.

W. E. Hipson, S. Kiritchenko, S. M. Mohammad, and R. J. Coplan. Examining the language of solitude versus loneliness in tweets. *Journal of Social and Personal Relationships*, 38 (5):1596–1610, May 2021. ISSN 0265-4075, 1460-3608. URL http://journals.sagepub.com/doi/10.1177/0265407521998460.

T. Hollenstein. This Time, It's Real: Affective Flexibility, Time Scales, Feedback Loops, and the Regulation of Emotion. *Emotion Review*, 7(4):308–315, Oct. 2015. ISSN 1754-0739, 1754-0747. URL http://journals.sagepub.com/doi/10.1177/1754073915590621.

M. Hosseinia, E. Dragut, and A. Mukherjee. Stance prediction for contemporary issues: Data and experiments. In L.-W. Ku and C.-T. Li, editors, *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 32–40, Online, July 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.socialnlp-1.5.

T. Hu, S. Wang, W. Luo, M. Zhang, X. Huang, Y. Yan, R. Liu, K. Ly, V. Kacker, B. She, and Z. Li. Revealing Public Opinion Towards COVID-19 Vaccines With Twitter Data in the United States: Spatiotemporal Perspective. *Journal of Medical Internet Research*, 23 (9):e30854, Sept. 2021. ISSN 1438-8871. URL https://www.jmir.org/2021/9/e30854.

B. Huang and K. Carley. Parameterized convolutional neural networks for aspect level sentiment classification. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1091–1096, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. URL https://aclanthology.org/D18-1136/.

J. Huang, M. Xiao, D. Li, Z. Jiang, Y. Yang, Y. Zhang, L. Qian, Y. Wang, X. Peng, Y. Ren, et al. Open-finllms: Open multimodal large language models for financial applications. *arXiv preprint arXiv:2408.11878*, 2024.

C. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225, 2014.

G. Juneja, G. Jajoo, N. Natarajan, H. Li, J. Jiao, and A. Sharma. Task facet learning: A structured approach to prompt optimization. *arXiv preprint arXiv:2406.10504*, 2024.

A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

J.-W. Kang and S.-Y. Choi. Comparative investigation of gpt and finbert's sentiment analysis performance in news across different sectors. *Electronics*, 14(6):1090, 2025.

A. Karanikola, G. Davrazos, C. M. Liapis, and S. Kotsiantis. Financial sentiment analysis: Classic methods vs. deep learning models. *Intelligent Decision Technologies*, 17(4):893–915, 2023. URL https://journals.sagepub.com/doi/abs/10.3233/IDT-230478.

C. Kearney and S. Liu. Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, 33:171–185, 2014.

O. Khattab, A. Singhvi, P. Maheshwari, Z. Zhang, K. Santhanam, S. Vardhamanan, S. Haq, A. Sharma, T. T. Joshi, H. Moazam, et al. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*, 2023.

T. Khot, H. Trivedi, M. Finlayson, Y. Fu, K. Richardson, P. Clark, and A. Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=_nGgzQjzaRy.

A. Kilgarriff. Word Senses. In E. Agirre and P. Edmonds, editors, *Word Sense Disambiguation: Algorithms and Applications*, pages 29–46. Springer Netherlands, Dordrecht, 2006. ISBN 978-1-4020-4809-8. URL https://doi.org/10.1007/978-1-4020-4809-8_2.

Y. Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. URL https://aclanthology.org/D14-1181.

J. Kobbe, I. Hulpuș, and H. Stuckenschmidt. Unsupervised stance detection for arguments from consequences. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 50–60, 2020.

A. Koch, A. Smith, S. T. Fiske, A. E. Abele, N. Ellemers, and V. Yzerbyt. Validating a brief measure of four facets of social evaluation. *Behavior Research Methods*, 56(8):8521–8539, 2024.

S. Kogan, D. Levin, B. R. Routledge, J. S. Sagi, and N. A. Smith. Predicting risk from financial reports with regression. In *Proceedings of human language technologies: the 2009 annual conference of the North American Chapter of the Association for Computational Linguistics*, pages 272–280, 2009.

T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

W. Kong, S. Hombaiah, M. Zhang, Q. Mei, and M. Bendersky. Prewrite: Prompt rewriting with reinforcement learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 594–601, 2024.

P. Koukaras, C. Nousi, and C. Tjortjis. Stock market prediction using microblogging sentiment analysis and machine learning. In *Telecom*, volume 3, pages 358–378. MDPI, 2022.

M. Kraus and S. Feuerriegel. Decision support from financial disclosures with deep neural networks and transfer learning. *Decision Support Systems*, 104:38–48, 2017.

T. Krone, C. J. Albers, P. Kuppens, and M. E. Timmerman. A multivariate statistical model for emotion dynamics. *Emotion*, 18(5):739–754, Aug. 2018. ISSN 1931-1516, 1528-3542. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/emo0000384.

D. Küçük and F. Can. Stance detection on tweets: An svm-based approach. *arXiv preprint arXiv:1803.08910*, 2018.

D. Küçük and F. Can. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53 (1):1–37, 2020.

S. Kumar, A. Y. Venkata, S. Khandelwal, B. Santra, P. Agrawal, and M. Gupta. SCULPT: Systematic tuning of long prompts. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14996–15029, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. URL https://aclanthology.org/2025.acl-long.730/.

P. Kuppens and P. Verduyn. Emotion dynamics. *Current Opinion in Psychology*, 17:22–26, Oct. 2017. ISSN 2352250X. URL https://linkinghub.elsevier.com/retrieve/pii/S2352250X16302019.

P. Kuppens, N. B. Allen, and L. B. Sheeber. Emotional Inertia and Psychological Maladjustment. *Psychological Science*, 21(7):984–991, July 2010a. ISSN 0956-7976, 1467-9280. URL http://journals.sagepub.com/doi/10.1177/0956797610372634.

P. Kuppens, Z. Oravecz, and F. Tuerlinckx. Feelings change: Accounting for individual differences in the temporal dynamics of affect. *Journal of Personality and Social Psychology*, 99(6):1042–1060, Dec. 2010b. ISSN 1939-1315, 0022-3514. URL https://doi.apa.org/doi/10.1037/a0020962.

M. Lai, V. Patti, G. Ruffo, and P. Rosso. Stance evolution and twitter interactions in an italian political debate. In *Natural Language Processing and Information Systems: 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings 23*, pages 15–27. Springer, 2018.

M. Lai, A. T. Cignarella, D. I. H. Farías, C. Bosco, V. Patti, and P. Rosso. Multilingual stance detection in social media political debates. *Computer Speech & Language*, 63: 101075, 2020.

S. Laine and T. Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

C. W. Leach, N. Ellemers, and M. Barreto. Group virtue: the importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of personality and social psychology*, 93(2):234, 2007.

W. Lee, J. Lee, and H. Kim. Logic: Llm-originated guidance for internal cognitive improvement of small language models in stance detection. *PeerJ Computer Science*, 10:e2585, 2024.

M. Levy, A. Jacoby, and Y. Goldberg. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15339–15353, 2024.

M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.acl-main.703/.

Q. Li, Z. Li, W. Liu, X. He, and Y. Pan. Sarcasm-gpt: advancing sarcasm detection with large language models. *The Computer Journal*, page bxaf055, 2025.

X. Li, S. Chan, X. Zhu, Y. Pei, Z. Ma, X. Liu, and S. Shah. Are chatgpt and gpt-4 general-purpose solvers for financial text analytics? a study on several typical tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 408–422, 2023a.

Y. Li and C. Caragea. Target-aware data augmentation for stance detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1850–1860, 2021.

Y. Li, T. Sosea, A. Sawant, A. J. Nair, D. Inkpen, and C. Caragea. P-stance: A large dataset for stance detection in political domain. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2355–2365, Online, Aug. 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.findings-acl.208/.

Y. Li, C. Zhao, and C. Caragea. Tts: A target-based teacher-student framework for zero-shot stance detection. In *Proceedings of the ACM Web Conference 2023*, pages 1500–1509, 2023b.

Y.-R. Lin and W.-T. Chung. The dynamics of twitter users' gun narratives across major mass shooting events. *Humanities and social sciences communications*, 7(1):1–16, 2020.

G. Lindelöf, T. Aledavood, and B. Keller. Dynamics of the Negative Discourse Toward COVID-19 Vaccines: Topic Modeling Study and an Annotated Data Set of Twitter Posts. *Journal of Medical Internet Research*, 25:e41319, Apr. 2023. ISSN 1438-8871. URL https://www.jmir.org/2023/1/e41319.

M. Liu and G. Shi. Enhancing llm-based text classification in political science: Automatic prompt optimization and dynamic exemplar selection for few-shot learning. *arXiv preprint arXiv:2409.01466*, 2024.

O. Liu, D. Fu, D. Yogatama, and W. Neiswanger. DeLLMa: Decision making under uncertainty with large language models. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=Acvo2RGSCy.

R. Liu, Z. Lin, Y. Tan, and W. Wang. Enhancing zero-shot and few-shot stance detection with commonsense knowledge graph. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3152–3157, 2021a.

R. Liu, J. Geng, A. J. Wu, I. Sucholutsky, T. Lombrozo, and T. L. Griffiths. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. *arXiv preprint arXiv:2410.21333*, 2024.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. URL https://arxiv.org/abs/1907.11692v1.

Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, et al. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, page 100017, 2023.

Y. Liu, J. Xu, L. L. Zhang, Q. Chen, X. Feng, Y. Chen, Z. Guo, Y. Yang, and P. Cheng.

Beyond prompt content: Enhancing llm performance via content-format integrated prompt optimization. *arXiv preprint arXiv:2502.04295*, 2025b.

Z. Liu, D. Huang, K. Huang, Z. Li, and J. Zhao. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, pages 4513–4519, 2021b.

T. Loughran and B. McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, 2011.

D. Loureiro, F. Barbieri, L. Neves, L. Espinosa Anke, and J. Camacho-collados. TimeLMs: Diachronic language models from Twitter. In V. Basile, Z. Kozareva, and S. Stajner, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland, May 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.acl-demo.25/.

N. Lozhnikov, L. Derczynski, and M. Mazzara. Stance prediction for russian: data and analysis. In *Proceedings of 6th International Conference in Software Engineering for Defence Applications: SEDA 2018 6*, pages 176–186. Springer, 2020.

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

Y. Luo, Z. Liu, Y. Shi, S. Z. Li, and Y. Zhang. Exploiting sentiment and common sense for zero-shot stance detection. In N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, and S.-H. Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*,

pages 7112–7123, Gyeongju, Republic of Korea, Oct. 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.621/.

J. C. Lyu, E. L. Han, and G. K. Luli. Covid-19 vaccine–related discussion on twitter: topic modeling and sentiment analysis. *Journal of medical Internet research*, 23(6):e24435, 2021.

J. Ma, C. Wang, H. Xing, D. Zhao, and Y. Zhang. Chain of stance: Stance detection with large language models. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 82–94. Springer, 2024.

A. Mamillapalli, B. Ogunleye, S. Timoteo Inacio, and O. Shobayo. Gruvader: Sentiment-informed stock market prediction. *Mathematics*, 12(23), 2024. ISSN 2227-7390. URL https://www.mdpi.com/2227-7390/12/23/3801.

G. J. McLachlan. Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. *Journal of the American Statistical Association*, 70(350):365–369, 1975.

Meta. Llama-3.3-70b-instruct. URL https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct. Accessed 2025-06-09.

M. Mets, A. Karjus, I. Ibrus, and M. Schich. Automated stance detection in complex topics and small languages: the challenging case of immigration in polarizing news media. *Plos one*, 19(4):e0302380, 2024.

L. Miao, M. Last, and M. Litvak. Twitter data augmentation for monitoring public opinion on covid-19 intervention measures. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, 2020.

MistralAI. Mistral small 3.1, 2025. URL https://mistral.ai/news/mistral-small-3-1. Accessed: 2025-06-30.

S. Mohammad. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. In *Proceedings of the 56th Annual Meeting of the Associa-*

tion for Computational Linguistics (Volume 1: Long Papers), pages 174–184, Melbourne, Australia, 2018. Association for Computational Linguistics. URL http://aclweb.org/anthology/P18-1017.

S. Mohammad and P. Turney. Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles, CA, June 2010. Association for Computational Linguistics. URL https://aclanthology.org/W10-0204.

S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41, 2016.

S. M. Mohammad. Words of warmth: Trust and sociability norms for over 26k English words. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18830–18850, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. URL https://aclanthology.org/2025.acl-long.922/.

S. M. Mohammad and P. D. Turney. CROWDSOURCING A WORD–EMOTION ASSOCIATION LEXICON. *Computational Intelligence*, 29(3):436–465, Aug. 2013. ISSN 0824-7935, 1467-8640. URL https://onlinelibrary.wiley.com/doi/10.1111/j.1467-8640.2012.00460.x.

S. M. Mohammad, P. Sobhani, and S. Kiritchenko. Stance and sentiment in tweets. *ACM Trans. Internet Technol.*, 17(3), June 2017. ISSN 1533-5399. URL https://doi.org/10.1145/3003433.

Y. Mu, M. Jin, C. Grimshaw, C. Scarton, K. Bontcheva, and X. Song. VaxxHesitancy: A Dataset for Studying Hesitancy towards COVID-19 Vaccination on Twitter. *Proceedings*

of the International AAAI Conference on Web and Social Media, 17:1052–1062, June 2023. ISSN 2334-0770, 2162-3449. URL https://ojs.aaai.org/index.php/ICWSM/article/view/22213.

R. Mukherjee, A. Bohra, A. Banerjee, S. Sharma, M. Hegde, A. Shaikh, S. Shrivastava, K. Dasgupta, N. Ganguly, S. Ghosh, et al. Ectsum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10893–10906, 2022.

M. P. Naeini, G. Cooper, and M. Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015. URL https://ojs.aaai.org/index.php/AAAI/article/view/9602.

Y. Nie, Y. Kong, X. Dong, J. M. Mulvey, H. V. Poor, Q. Wen, and S. Zohren. A survey of large language models for financial applications: Progress, prospects and challenges. *arXiv preprint arXiv:2406.11903*, 2024.

F. Niu, M. Yang, A. Li, B. Zhang, X. Peng, and B. Zhang. A challenge dataset and effective models for conversational stance detection. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 122–132, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.11/.

N. N. Oosterhof and A. Todorov. The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, 105(32):11087–11092, 2008.

OpenAI. Introducing openai o3 and o4-mini, April 2025. URL https://openai.com/index/introducing-o3-and-o4-mini/. Accessed: 2025-06-03.

OpenAI. Introducing gpt-4.1 in the api, Apr. 2025. URL https://openai.com/index/gpt-4-1/. Accessed 9 Jun 2025.

OpenAI. Gpt-5 system card. https://cdn.openai.com/gpt-5-system-card.pdf, Aug. 2025.

OpenAI. Model release notes – introducing gpt-4.1 mini, May 2025. URL https://help.openai.com/en/articles/9624314-model-release-notes. Accessed 9 Jun 2025.

OpenAI. Introducing gpt-oss, Aug. 2025. URL https://openai.com/index/introducing-gpt-oss/. Accessed: 2025-08-17.

V. C. Osuji, E. M. Galante, D. Mischoulon, J. E. Slaven, and G. Maupome. Covid-19 vaccine: A 2021 analysis of perceptions on vaccine safety and promise in a us sample. *PLoS One*, 17(5):e0268784, 2022.

J. Pang, J. Wei, A. P. Shah, Z. Zhu, Y. Wang, C. Qian, Y. Liu, Y. Bao, and W. Wei. Improving data efficiency via curating LLM-driven rating systems. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=DKkQtRMowq.

L. Pangtey, A. Bhatnagar, S. Bansal, S. S. Dar, and N. Kumar. Large language models meet stance detection: A survey of tasks, methods, applications, challenges and future directions. *arXiv preprint arXiv:2505.08464*, 2025.

B. Peng, E. Chersoni, Y.-Y. Hsu, and C.-R. Huang. Is domain adaptation worth your investment? comparing bert and finbert on financial tasks. In *Proceedings of the Third Workshop on Economics and Natural Language Processing*, pages 37–44, 2021.

G. Pleiss, T. Zhang, E. Elenberg, and K. Q. Weinberger. Identifying mislabeled data using the area under the margin ranking. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17044–17056. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c6102b3727b2a7d8b1bb6981147081ef-Paper.pdf.

R. Plutchik. *The emotions*. University Press of America, 1991.

S. Poddar, R. Mukherjee, S. Khatuya, N. Ganguly, and S. Ghosh. How covid-19 has impacted the anti-vaccine discourse: A large-scale twitter study spanning pre-covid and post-covid era. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 1276–1288, 2024.

S. Praveen, R. Ittamalla, and G. Deepak. Analyzing the attitude of Indian citizens towards COVID-19 vaccine – A text analytics study. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 15(2):595–599, Mar. 2021. ISSN 18714021. URL https://linkinghub.elsevier.com/retrieve/pii/S1871402121000618.

R. Pryzant, D. Iter, J. Li, Y. Lee, C. Zhu, and M. Zeng. Automatic prompt optimization with "gradient descent" and beam search. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7957–7968, 2023.

M. Qorib, T. Oladunni, M. Denis, E. Ososanya, and P. Cotae. Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. *Expert Systems with Applications*, 212:118715, Feb. 2023. ISSN 09574174. URL https://linkinghub.elsevier.com/retrieve/pii/S0957417422017407.

L. Ranaldi and A. Freitas. Aligning large and small language models via chain-of-thought reasoning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1812–1827, 2024.

R. Rappuoli, C. W. Mandl, S. Black, and E. De Gregorio. Vaccines for the twenty-first century society. *Nature Reviews Immunology*, 11(12):865–872, Dec. 2011. ISSN 1474-1733, 1474-1741. URL https://www.nature.com/articles/nri3085.

N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.

M. Renze. The effect of sampling temperature on problem solving in large language models. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7346–7356, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.432. URL https://aclanthology.org/2024.findings-emnlp.432/.

G. Roussos and Y. Dunham. The development of stereotype content: The use of warmth and competence in assessing social groups. *Journal of Experimental Child Psychology*, 141: 133–144, 2016.

S. Rude, E.-M. Gortner, and J. Pennebaker. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133, Dec. 2004. ISSN 0269-9931, 1464-0600. URL http://www.tandfonline.com/doi/abs/10.1080/02699930441000030.

M. Sajjadi, M. Javanmardi, and T. Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29:1163–1171, 2016.

F. Sakketou, A. Lahnala, L. Vogel, and L. Flek. Investigating user radicalization: A novel dataset for identifying fine-grained temporal shifts in opinion. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3798–3808, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.405/.

S. N. Saleh, S. A. McDonald, M. A. Basit, S. Kumar, R. J. Arasaratnam, T. M. Perl, C. U. Lehmann, and R. J. Medford. Public perception of COVID-19 vaccines through analysis of Twitter content and users. *Vaccine*, 41(33):4844–4853, July 2023. ISSN 0264410X. URL https://linkinghub.elsevier.com/retrieve/pii/S0264410X23007430.

A. Sarirete. Sentiment analysis tracking of COVID-19 vaccine through tweets. *Journal of Ambient Intelligence and Humanized Computing*, 14(11):14661–14669, Nov. 2023. ISSN 1868-5137, 1868-5145. URL https://link.springer.com/10.1007/s12652-022-03805-0.

T. Schnabel and J. Neville. Prompts as programs: A structure-aware approach to efficient compile-time prompt optimization. *arXiv e-prints*, pages arXiv–2404, 2024.

S. Schulhoff, M. Ilie, N. Balepur, K. Kahadze, A. Liu, C. Si, Y. Li, A. Gupta, H. Han, S. Schulhoff, et al. The prompt report: a systematic survey of prompt engineering techniques. *arXiv preprint arXiv:2406.06608*, 2024.

R. P. Schumaker and H. Chen. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems (TOIS)*, 27(2):1–19, 2009.

M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997. URL https://www.doi.org/10.1109/78.650093.

Z. Shi, J. Wei, Z. Xu, and Y. Liang. Why larger language models do in-context learning differently? In *Proceedings of the 41st International Conference on Machine Learning*, pages 44991–45013, 2024.

S. Shin and Y. Kim. Enhancing graph of thought: Enhancing prompts with llm rationales and dynamic temperature control. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

U. A. Siddiqua, A. N. Chy, and M. Aono. Tweet stance detection using an attention based neural ensemble model. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)*, pages 1868–1873, 2019.

V. K. Singh, P. Mohankumar, and A. Kamal. Fin-stance: A novel deep learning-based multi-task model for detecting financial stance and sentiment. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE, 2023.

I. Sirbu, T. Sosea, C. Caragea, D. Caragea, and T. Rebedea. Multimodal semi-supervised learning for disaster tweet classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2711–2723, 2022.

V. Slovikovskaya and G. Attardi. Transfer learning from transformers to fake news challenge stance detection (fnc-1) task. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1211–1218, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.152/.

P. Sobhani, D. Inkpen, and X. Zhu. A dataset for multi-target stance detection. In M. Lapata, P. Blunsom, and A. Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557, Valencia, Spain, Apr. 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-2088.

S. Sohangir, D. Wang, A. Pomeranets, and T. M. Khoshgoftaar. Big data: Deep learning for financial sentiment analysis. *Journal of Big Data*, 5(1):1–25, 2018.

K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

G. H. Soong and C. C. Tan. Sentiment analysis on 10-k financial reports using machine

learning approaches. In *2021 IEEE 11th International Conference on System Engineering and Technology (ICSET)*, pages 124–129. IEEE, 2021.

T. Sosea and C. Caragea. Leveraging training dynamics and self-training for text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4750–4762, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.350.

C. Stab, T. Miller, B. Schiller, P. Rai, and I. Gurevych. Cross-topic argument mining from heterogeneous sources. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. URL https://aclanthology.org/D18-1402/.

M. Stella, V. Restocchi, and S. De Deyne. #lockdown: Network-Enhanced Emotional Profiling in the Time of COVID-19. *Big Data and Cognitive Computing*, 4(2):14, June 2020. ISSN 2504-2289. URL https://www.mdpi.com/2504-2289/4/2/14.

Q. Sun, Z. Wang, Q. Zhu, and G. Zhou. Stance detection with hierarchical attention network. In *Proceedings of the 27th international conference on computational linguistics*, pages 2399–2409, 2018.

Q. Sun, Z. Wang, S. Li, Q. Zhu, and G. Zhou. Stance detection via sentiment information and neural network model. *Frontiers of Computer Science*, 13:127–138, 2019.

J. K. Swencionis, C. H. Dupree, and S. T. Fiske. Warmth-competence tradeoffs in impression management across race and social-class divides. *Journal of Social Issues*, 73(1):175–191, 2017.

X. Tang, X. Wang, W. X. Zhao, S. Lu, Y. Li, and J.-R. Wen. Unleashing the potential of large language models as prompt optimizers: Analogical analysis with gradient-based model optimizers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25264–25272, 2025.

D. Teodorescu and S. Mohammad. Evaluating emotion arcs across languages: Bridging the global divide in sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4124–4137, 2023.

P. Trivedi, S. Chakraborty, A. Reddy, V. Aggarwal, A. S. Bedi, and G. K. Atia. Alignpro: A principled approach to prompt optimization for llm alignment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(26):27653–27661, Apr. 2025. URL https://ojs.aaai.org/index.php/AAAI/article/view/34979.

J. Vamvas and R. Sennrich. X-stance: A multilingual multi-target dataset for stance detection. *arXiv preprint arXiv:2003.08385*, 2020.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

R. Villa-Cox, S. Kumar, M. Babcock, and K. M. Carley. Stance in replies and quotes (srq): A new dataset for learning stance in twitter conversations. *arXiv preprint arXiv:2006.00691*, 2020.

K. Vishnubhotla and S. M. Mohammad. Tweet Emotion Dynamics: Emotion word usage in tweets from US and Canada. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4162–4176, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.442/.

X. Wan, H. Zhou, R. Sun, H. Nakhost, K. Jiang, and S. Ö. Arık. From few to many: Self-improving many-shot reasoners through iterative optimization and generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=JBXO05r4AV.

S. Wang, Z. Chen, C. Shi, C. Shen, and J. Li. Mixture of demonstrations for in-context learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan,

U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 88091–88116. Curran Associates, Inc., 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/a0da098e0031f58269efdcba40eedf47-Paper-Conference.pdf.

W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

X. Wang and M. Brorsson. Can large language model analyze financial statements well? In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 196–206, 2025.

X. Wang, C. Li, Z. Wang, F. Bai, H. Luo, J. Zhang, N. Jojic, E. P. Xing, and Z. Hu. Promptagent: Strategic planning with language models enables expert-level prompt optimization. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=22pyNMuIoa.

J. Wei and K. Zou. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. URL https://aclanthology.org/D19-1670/.

J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=gEZrGCozdqR.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022b.

X. Wei and L. Liu. Are large language models good in-context learners for financial sentiment analysis? *arXiv preprint arXiv:2503.04873*, 2025.

A. Williams, N. Nangia, and S. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In M. Walker, H. Ji, and A. Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. URL https://aclanthology.org/N18-1101/.

B. Wojciszke, A. E. Abele, and W. Baryla. Two dimensions of interpersonal attitudes: Liking depends on communion, respect depends on agency. *European Journal of Social Psychology*, 39(6):973–990, 2009.

O. T. Wu, F. K. S. Chan, Z. Zhang, Y. N. Law, B. Drescher, and E. S. B. Lai. Prompt selection and augmentation for few examples code generation in large language model and its application in robotics control. *arXiv preprint arXiv:2403.12999*, 2024.

Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.

C. Xu, C. Paris, S. Nepal, R. Sparks, C. Long, and Y. Wang. Dan: Dual-view representation learning for adapting stance classifiers to new domains. *arXiv preprint arXiv:2003.06514*, 2020.

H. Xu, S. Vucetic, and W. Yin. OpenStance: Real-world zero-shot stance detection. In A. Fokkens and V. Srikumar, editors, *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 314–324, Abu Dhabi, United

Arab Emirates (Hybrid), Dec. 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.conll-1.21/.

W. Xue and T. Li. Aspect based sentiment analysis with gated convolutional networks. In I. Gurevych and Y. Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2514–2523, Melbourne, Australia, July 2018. Association for Computational Linguistics. URL https://aclanthology.org/P18-1234/.

H. Yakura. Evaluating large language models' ability using a psychiatric screening tool based on metaphor and sarcasm scenarios. *Journal of Intelligence*, 12(7):70, 2024.

C. Yang, X. Wang, Y. Lu, H. Liu, Q. V. Le, D. Zhou, and X. Chen. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bb4VGOWELI.

S. Yousefinaghani, R. Dara, S. Mubareka, A. Papadopoulos, and S. Sharif. An analysis of COVID-19 vaccine sentiments and opinions on Twitter. *International Journal of Infectious Diseases*, 108:256–262, July 2021. ISSN 12019712. URL https://linkinghub.elsevier.com/retrieve/pii/S1201971221004628.

G. Zarrella and A. Marsh. MITRE at SemEval-2016 task 6: Transfer learning for stance detection. In S. Bethard, M. Carpuat, D. Cer, D. Jurgens, P. Nakov, and T. Zesch, editors, *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 458–463, San Diego, California, June 2016. Association for Computational Linguistics. URL https://aclanthology.org/S16-1074/.

W. Zeng, C. Xu, Y. Zhao, J.-G. Lou, and W. Chen. Automatic instruction evolving for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6998–7018, 2024.

B. Zhang, D. Ding, and L. Jing. How would stance detection techniques evolve after the launch of chatgpt? *arXiv preprint arXiv:2212.14548*, 2022.

B. Zhang, H. Yang, and X.-Y. Liu. Instruct-FinGPT: Financial Sentiment Analysis by Instruction Tuning of General-Purpose Large Language Models. FinLLM Workshop at IJCAI 2023, June 2023. URL https://ssrn.com/abstract=4489831. Available at SSRN.

H. Zhang, H. Zhang, Z. Liu, B. Gu, and Y. Chang. Collaborative discrete-continuous black-box prompt learning for language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

Y. Zhou, A. I. Cristea, and L. Shi. Connecting targets to tweets: Semantic attention-based model for target-specific stance detection. In *Web Information Systems Engineering–WISE 2017: 18th International Conference, Puschino, Russia, October 7-11, 2017, Proceedings, Part I 18*, pages 18–32. Springer, 2017.

Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*, 2022.

Y. Zhu, Z. Yin, G. Tyson, E.-U. Haq, L.-H. Lee, and P. Hui. Apt-pipe: A prompt-tuning tool for social data annotation using chatgpt. In *Proceedings of the ACM Web Conference 2024*, pages 245–255, 2024.

E. Zotova, R. Agerri, M. Nuñez, and G. Rigau. Multilingual stance detection in tweets: The catalonia independence corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1368–1375, 2020.

Appendix A

Appendix: GunStance: Stance detection for gun control and gun regulation

A.1 Existing stance detection datasets

Table A1 shows the comparison of various Target-specific Stance detection datasets.

A.2 Crawling rules/terms

We use a set of seed rules[1] as shown in Table A2 to initiate the crawling of tweets for each event. Table A3 contains the additional terms we added to the rule based on frequent hashtags. In collecting these additional terms, we filtered out the terms containing *news* as a substring and terms like *mass, dead, breaking, update, usa, america* to reduce noise in the data.

[1] https://developer.twitter.com/en/docs/twitter-api/tweets/search/integrate/build-a-query

Table A1: Summary of publicly available stance-detection datasets.

Dataset	Source	# Targets	Targets	Size
SemEval-16 (Mohammad et al., 2016)	Twitter	6	Atheism; Climate change is a real concern; Feminist Movement; Hillary Clinton; Legalization of Abortion; Donald Trump	48,700
MultiTarget (Sobhani et al., 2017)	Twitter	4	Clinton–Sanders; Clinton–Trump; Cruz–Trump	4,455
Cross-Topic AM (Stab et al., 2018)	News, blogs, forums	8	Abortion; Cloning; Death penalty; Gun control; Marijuana legalization; Minimum wage; Nuclear energy; School uniforms	25,492
WTWT (Conforti et al., 2020b)	Twitter	5	Mergers: Cigna–Express Scripts; Aetna–Humana; CVS–Aetna; Anthem–Cigna; Disney–Fox	51,284
SRQ (Villa-Cox et al., 2020)	Twitter	4	General terms; Iran Deal; Santa Fe; Shooting Student Marches	5,221
VAST (Allaway and McKeown, 2020)	NYT comments	5,634	Targets from *Room for Debate* (ARC corpus)	18,545
ProCon-20 (Hosseinia et al., 2020)	procon.org	419	Controversial issues and responses	6,094
StArCon (Kobbe et al., 2020)	Debatepedia	190	Various topics	5,398
COVID-19 (Glandt et al., 2021)	Twitter	4	Anthony Fauci; Keeping schools closed; Stay-at-home orders; Wearing a face mask	6,133
P-Stance (Li et al., 2021)	Twitter	3	Donald Trump; Joe Biden; Bernie Sanders	21,574
StEduCov (Hamad et al., 2022)	Twitter	1	Online education during COVID-19	16,572
MT-CSD (Niu et al., 2024)	Reddit	5	Bitcoin; Tesla; SpaceX; Joe Biden; Donald Trump	15,876
GunStance (Ours)	Twitter	2	Banning guns; Regulating guns	5,500

Table A2: Seed rules used to crawl tweets for each event.

Location	**Seed rules**
Atlanta, Georgia	*((atlanta OR ackworth OR georgia OR atlantaga OR ackworthga) (shooting OR massacre))*
Boulder, Colorado	*((boulder OR colorado) (shooting OR massacre OR strong)) OR #bouldershooting OR #BoulderMassacre OR #BoulderStrong*
San Jose, California	*(((san jose) OR sanjose OR (santa clara) OR (santaclara) OR vta) (shooting OR #californiashooting))*
Buffalo, New York	*((buffalo OR ny OR (new york) OR newyork) shooting) OR buffaloshooting*
Uvalde, Texas	*((uvalde OR texas) shooting) OR uvaldeshooting OR uvaldetexas*
Tulsa, Oklahoma	*((tulsa OR tulsaok OR oklahoma OR tulsaoklahoma) shooting) OR tulsashooting*
Highland Park, Illinois	*((highlandpark OR (highland park) OR illinois OR highlandparkillinois) shooting) OR highlandparkshooting*

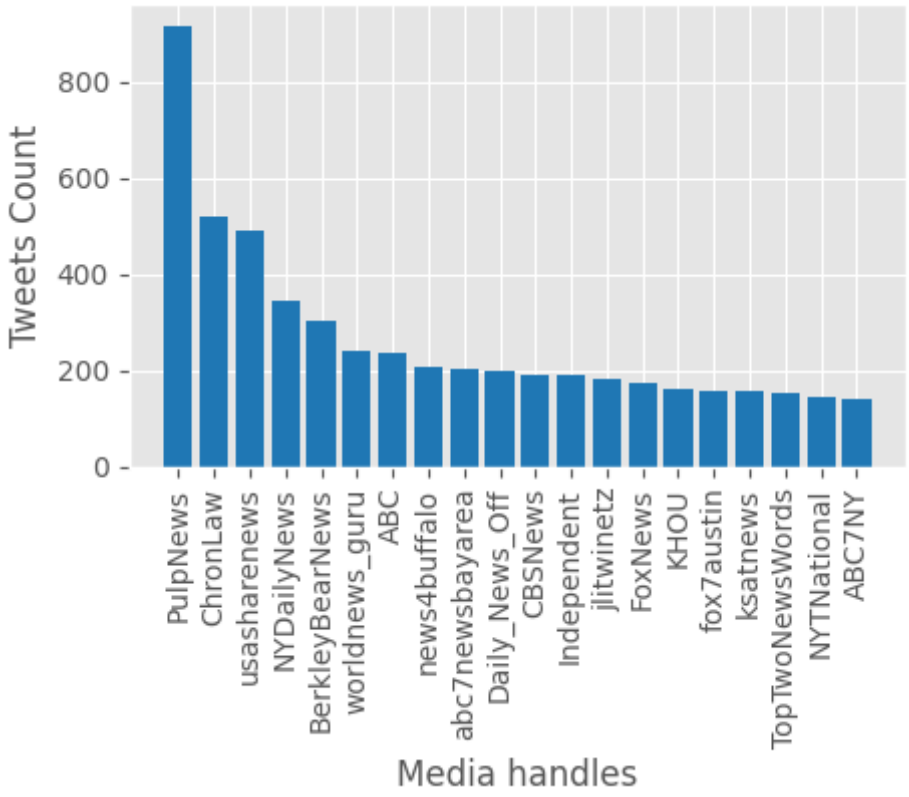


Figure A.2.1: Plot of the top 20 news media outlets ordered by the number of tweets posted.

Table A3: Additional terms added to the rule based on frequent hashtags.

Location	**Week 1**	**Weeks 2–7**
Atlanta, Georgia	asian, massageparlor, activeshooter, woodstock, stopasianhate, stopaapihate, asianlivesmatter, atlantaspa, stopasianhatecrimes, aapi, hatecrime, racism, asianamericans, atlantamassacre	boulder, thalapathy65, master, thalapathyvijay
Boulder, Colorado	bouldercolorado, kingsoopers, guncontrolnow, activeshooter, coloradoshooting, police, zfgvideography, gunreform, gunsense, guncontrol, gunreformnow, gunviolence, atlantamassacre, nra	banassaultweapons, copolitics
San Jose, California	california, gunreformnow, guncontrolnow, massshooting, sanjosestrong, gunviolence, sanjoseshooting, guncontrol, gunsense, jose, transit, transportation, sf, sanfrancisco	sfmta, bart, muni
Buffalo, New York	buffalony, buffalostrong, massshooting, guncontrolnow, racism, hatecrime, paytongendron, whitesupremacist, whitesupremacy, buffalonewyork, supermarket, buffalomassacre, blacklivesmatter	uvalde, ushistory, texas, nra, endthefilibuster, gunviolence
Uvalde, Texas	guncontrolnow, guncontrol, robbelementaryschool, gunviolence, uvaldetx, uvaldemassacre, nra, robbelementary, texasschoolmassacre, uvaldeschoolmassacre	specialsessionnow, uvaldepolice, gunreformnow, endgunviolence, banassaultweaponsnow, enough, dosomething, uvaldecoverup, uvaldepolicecowards
Tulsa, Oklahoma	guncontrolnow, gunreformnow, guncontrol, uvalde, massshooting, enoughisenough, tulsamassacre, hospital, tulsahospital, care2, gunviolence	-
Highland Park, Illinois	chicago, highlandparkparade, 4thofjuly, highland, massshooting, guncontrolnow, robertcrimo, guncontrol, highlandparkmassacre, prolifemyass, ushistory	park

A.3 Tweets statistics

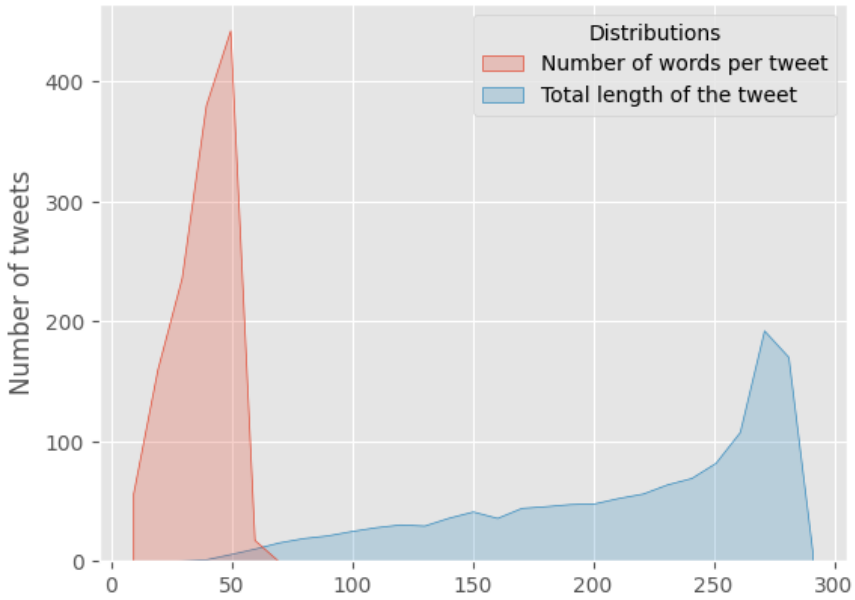


Figure A.3.2: Distributions for the number of words per tweet (red) and total character length of the tweet (blue). The mode of the word distribution is 47 words, while the mode of the length distribution is 278 characters.

A.4 Baseline models

Supervised baseline models

1. **BiLSTM:** We utilize Bi-Directional Long Short-Term Memory Networks (BiLSTMs) (Schuster and Paliwal, 1997) as a supervised model. The model takes tweets as input and is trained to predict the stance toward a target without explicitly incorporating the target information.

2. **Kim-CNN:** We employ Convolutional Neural Networks (CNNs) for text, proposed by (Kim, 2014), as another supervised model. Similar to BiLSTM, the CNN model takes tweets as input and is trained to predict the stance towards a target without directly incorporating target information.

3. **PGCNN:** We utilize parameterized Convolutional Neural Networks proposed by (Huang

and Carley, 2018). PGCNN utilizes parameterized filters and parameterized gates to capture the aspect specific features for sentiment classification.

4. **TAN:** We utilize Target-specific Attention Networks (Du et al., 2017), an attention-based BiLSTM model. As opposed to the BiLSTM and Kim-CNN models, TAN identifies features specific to the target of interest by explicitly incorporating the target information.

5. **ATGRU:** The Bi-Directional Gated Recurrent Unit (GRU) Network with Token-Level Attention Mechanism (Zhou et al., 2017) is an attention-based Bi-GRU model that also explicitly uses target information. It identifies specific target features using the attention mechanism.

6. **GCAE:** The Gated Convolutional Network with Aspect Embedding (Xue and Li, 2018) is a CNN-based model which, in addition to tweets, incorporates target information and also employs a gating mechanism to filter out information that is not related to the target.

7. **BERTweet:** We employ a pre-trained Transformer based model from (Loureiro et al., 2022). This model is based on RoBERTa (Liu et al., 2019) and is pre-trained on a large corpus of tweets. We fine-tuned the model using the labeled data.

Note: We have chosen to use well-established supervised methods for stance detection in our comparisons, as our focus was on zero-shot and semi-supervised methods, including a novel SSL/LLM hybrid model. While more recent, less established supervised models may perform better than those we have used in our comparisons, the current trend in the literature is to focus on models that work with a small amount of data (if any), while leveraging pre-trained models and/or unlabeled data (e.g., zero-shot, few-shot, semi-supervised). We have followed this trend since the zero-shot and semi-supervised scenarios fit very well with the dataset that we assembled.

A.4.1 Semi-supervised baselines

Given the availability of a significant amount of unlabeled data for each target in the dataset, we also investigate semi-supervised models capable of leveraging such unlabeled data. The following semi-supervised baseline models were employed:

1. **BERT-NS:** Similar to Glandt et al. (2021), we employ the self-training with Noisy Student (NS) method (Xie et al., 2020), a semi-supervised learning approach that leverages self-training and knowledge distillation (Hinton et al., 2015) to enhance the performance of a teacher model using unlabeled data. Initially, a teacher model is trained using the available labeled data and then utilized to generate pseudo-labels for the unlabeled data. Subsequently, a noisy student model is trained using both the labeled and pseudo-labeled data. This process can be iterated multiple times by replacing the teacher with the student. In our study, both the teacher and the student models represent fine-tuned BERTweet models.

2. **FixMatch:** Introduced by Sohn et al. (2020), FixMatch combines two common SSL techniques: consistency regularization (Laine and Aila, 2016; Sajjadi et al., 2016) and pseudo-labeling (McLachlan, 1975). As part of the FixMatch procedure, two augmented versions are created for the unlabeled examples: a weakly augmented version of an unlabeled example is passed through the model and the prediction is converted into a pseudo-label; then, the model is trained to predict that pseudo-label when fed with a strongly augmented version of the same unlabeled example. While originally designed for image classification, FixMatch has also shown good performance when used on text data in a multimodal setup (Sirbu et al., 2022). Hence, we also employ FixMatch as a strong semi-supervised baseline.

3. **AUM-ST:** AUM-ST is a novel SSL method that takes advantage of the training dynamics of unlabeled data to enhance a model's performance. This framework extends the concept of Area Under the Margin (Pleiss et al., 2020) to unlabeled data to identify potentially incorrect pseudo-labels. The authors show that removing examples with low

AUM scores and using data augmentations such as synonym replacement, switchout, and back-translations can significantly boost the generalization performance.

A.4.2 Zero-shot baseline models

1. **BART-NLI:** Natural Language Inference (NLI) is the task of determining whether a given hypothesis entails, contradicts, or is neutral to a given premise. As shown by Li et al. (2023b), the BART (Lewis et al., 2020) model pre-trained on the MultiNLI dataset (Williams et al., 2018) can be used for zero-shot stance detection by creating a semantic mapping between the stance labels and the NLI labels. Thus, we consider BART-NLI as a strong zero-shot baseline for our dataset.

2. **FLAN:** The main idea is to formulate the task in such a way that the language model generates the answer by completing the text. FLAN (Wei et al., 2022a) improves the zero-shot learning capabilities of language models by introducing *instruction tuning*, and we consider it as another strong zero-shot baseline in our experiments.

3. **ChatGPT:** One of the most notable large language models (LLMs) that has been receiving increasing interest from the research community since its release in December 2022 is ChatGPT, as it demonstrated impressive performances in various domains (e.g. healthcare, education, scientific research, etc.) and on diverse downstream NLP tasks (e.g. question answering, text classification, text generation, code generation, etc.)(Liu et al., 2023). In particular, (Zhang et al., 2022) shows that ChatGPT obtains state-of-the-art results on commonly used stance detection datasets such as SemEval-2016 (Mohammad et al., 2016) and P-Stance (Li et al., 2021), so we consider it as another strong zero-shot baseline for our dataset.

A.5 Experimental setup

Hyperparameters

To determine suitable hyperparameters for the models, the development set was utilized. We utilized the Adam optimizer for all the supervised and semi-supervised models. However, for the supervised model, excluding BERTweet, we used a learning rate of 1e-5, weight decay of 4e-5, gradient clipping with the maximum norm of 4.0, and a dropout rate of 0.5 to the linear classification layer. The training process of those models (except for BERTweet) involved 120 epochs, with a mini-batch size of 32 for each iteration. For all models, we performed 3 runs (starting with different random seeds) and report average scores. Further details regarding the specific hyperparameters for each model are presented below:

BiLSTM, ATGRU, TAN: Each of these Recurrent Neural Network models used a hidden unit with 512 dimensions and a dropout of 0.2.

GCAE, Kim-CNN: These CNN-based models utilized filters of width 2, 3, 4, and 5. Each filter width was associated with 25 feature maps. After the convolutional layers, a linear classifier with a hidden dimension of 128 was employed.

BERTweet: This model was trained with a learning rate of 2e-5 using a linear scheduler with warmup steps equivalent to 10% of the total training steps. A batch size of 32 was employed, and the model was trained for 100 epochs. Early stopping criteria based on the F1-score were utilized with patience of 10 epochs. The model was trained using the Adam optimizer and cross-entropy loss.

BERT-NS: For BERT-NS, both teacher and student models were trained with hyperparameters similar to BERTweet. The confidence threshold for selecting pseudo-labels was set to 0.95.

AUM-ST: We utilized a supervised batch size of 32 and an unsupervised batch size of 128. To filter unlabeled data, we applied a threshold of 0.7 and an augmentation percentile of 0.5. The self-training process consisted of 50 steps. As augmentation strengths, we employed a mix weak augmentation strength of 0 and a max weak augmentation strength of 2. For strong augmentation, we utilized a min strength of 3 and a max strength of 40.

FixMatch: We utilized the BERTweet model in FixMatch. For the FixMatch-specific parameters, we used an unlabeled loss weight $\lambda_u = 1$, a ratio between unlabeled and labeled examples $\mu = 7$, and a threshold $\tau = 0.7$ for using only the pseudo-labels with high confidence. For the text augmentation, we used the same techniques and strengths employed for AUM-ST.

Prompt engineering

In the case of the zero-shot models employed, it is important to pay close attention to the design of the prompts used to obtain predictions.

BART-NLI: Similar to Li et al. (2023b), we consider a mapping between predicting stance labels (*Against*, *In-Favor*, *Neutral*) and the task of predicting entailment labels (*Contradiction*, *Entailment*, *Neutral*), by using the tweet as a premise and designing a prompt that contains the target for the hypothesis. We experimented with several prompt templates and chose "I think [target]!" as the best one, where [target] is either "guns should be banned" or "guns should be regulated." We use the BART-large[2] version of the model.

FLAN: We experimented with several prompt templates and chose "[tweet] What is the stance of the writer? (A) [neg_target] (B) [target] (C) neutral" as the best performing one, where [target] is either "guns should be banned" or "guns should be regulated" and [neg_target] is the corresponding "guns should not be banned" or "guns should not be regulated." We use the FLAN-T5 XXL[3] model.

[2] https://huggingface.co/facebook/bart-large-mnli
[3] https://huggingface.co/google/flan-t5-xxl

ChatGPT: We use the Chat Completions API [4] provided by OpenAI. For the system message that sets the behaviour of the assistant we use the instruction 'You will be provided with a tweet, and your task is to classify its stance as "[target]", "neutral", or "[neg_target]".' Then, for the user message we provide the request '[tweet]'. The ChatGPT version used is 'gpt-3.5-turbo-0613'. Because the answer of the model may come in different forms (e.g. "guns should be banned", "Stance: guns should be banned"), the final label is considered the one stance ("[target]", "neutral", or "[neg_target]") that is a substring of the answer given by the model. When this is true for neither of them (e.g. "I'm sorry, but I can't assist with that."), the final label for evaluation purposes is considered to be "neutral". In the streaming pipeline defined by AUM-ST$_{+LLM}$, the case where zero or more than one stances are substrings of the answer, We consider ChatGPT has provided an **inconclusive** label, so that tweet won't be used. This happens for only about 0.5% of the cases.

Proposed hybrid SSL/LLM approach

For AUM-ST$_{+ChatGPT}$, we use the same setups as AUM-ST and we generate the ChatGPT pseudo-labels in the same manner as the zero-shot baseline. Additionally, at each self-training iteration we select around 5% of unlabeled examples (k in Step 4 of the algorithm) with AUMs close to zero to be passed through ChatGPT.

Evaluation metrics

In order to evaluate the supervised, semi-supervised and zero-shot models, we employ commonly used metrics for classification tasks: Accuracy, Precision, Recall, and F1 score. We denote true positives, true negatives, false positives, and false negatives by TP, TN, FP, and FN, respectively.

Accuracy measures the overall correctness of the model's predictions by calculating the ratio of correctly predicted instances to the total number of instances. It is calculated as:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{A.1}$$

[4] https://platform.openai.com/docs/guides/text-generation/chat-completions-api

Precision measures the proportion of correctly predicted positive instances out of all the instances that were predicted as positive. It is calculated as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{A.2}$$

Recall measures the proportion of correctly predicted positive instances out of all the actual positive instances and is calculated as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{A.3}$$

F1-Score is a commonly used metric that combines precision and recall into a single metric, providing a trade-off between these two metrics. The F1-score is calculated as:

$$\text{F1-score} = \frac{2 * TP}{2 * TP + FP + FN} \tag{A.4}$$

We also utilized the **Expected Calibration Error (ECE)** (Naeini et al., 2015) metric to measure how reliably a model's predicted probabilities reflect the true probabilities. ECE quantifies the difference between the predicted confidence of a model and its accuracy. A well-calibrated model should have a low ECE, implying that the model's predicted probabilities more accurately reflect the true probabilities. The ECE is calculated as:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)| \tag{A.5}$$

where, n is the number of samples, M is the equally-spaced bins, B_m is the set of indices of samples whose prediction confidence falls into the respective bins, and $\text{acc}(B_m)$ and $\text{conf}(B_m)$ represent the accuracy and confidence within the bin B_m, respectively (Guo et al., 2017).

Appendix B

Appendix: The shifting landscape of vaccine discourse: A massive dataset examining a decade of social media posts on vaccines before and during the COVID-19 pandemic

B.1 Prompt for Llama 3.3 model

To generate the stance of the author towards vaccines in a post, we used the following approach with the Llama 3.3-70B Instruction-tuned model. The process involved sending a series of messages to the model, which included:

```
messages = [
{"role": "system", "content": "You are a very good stance
   detection model that can accurately detect the author's
   stance towards vaccines in a post."},
```

```
{"role": "user", "content":  """# Task: Classify the stance of
   a tweet towards the target 'vaccines' as either favor,
   against, or neutral.
# Guidelines:
a) Main Topic: Determine if the tweet is primarily about human
   vaccines. If the tweet contains vaccine keywords but the
   main topic is not vaccines, classify it as neutral.
b) Stance Indicators: Look for keywords, phrases, emoticons,
   and tone that indicate a Favor, Against, or Neutral attitude
    towards vaccines.
c) Implicit vs. Explicit Stance: Consider both implicit and
   explicit stance expressions. If a tweet implies a stance
   without directly stating it, use the context and language to
    determine the intended stance.
d) Sarcasm and Irony: Be aware of sarcasm and irony, which can
   be used to express an opposing stance. Look for
   inconsistencies between the literal meaning and the tone or
   language used.

## Favor: Classify the tweet as Favor if it:
1. Shows an explicitly positive stance towards vaccines: openly
    supports vaccination or urges others to get vaccinated
2. Uses pro-vaccine keywords: languages that frame vaccines as
   beneficial, life-saving, or socially responsible
3. Highlights safety or effectiveness: praises scientific
   research, high efficacy rates, or robust safety records
```

4. Praises vaccination efforts: encourages vaccination, praises vaccine efficacy, or expresses gratitude for vaccines

Against: Classify the tweet as against if it:
1. Shows an explicitly negative stance towards vaccines: openly opposes vaccination or urges others not to vaccinate
2. Uses anti-vaccine keywords: languages that frame vaccines as harmful, unnecessary, or deceitful
3. Attacks safety or effectiveness: alleges side-effects, conspiracies, or inadequate testing
4. Criticizes vaccination efforts or advocates: mocks mandates, scientists, health agencies, or pro-vaccine campaigns

Neutral: Classify the tweet as neutral if it:
1. Shows no clear stance towards vaccines: it neither supports nor opposes vaccination
2. Uses neutral or tentative language: e.g., simply asking a question or stating an opinion that signals indifference/ uncertainty.
3. Presents facts or information only: shares data, news, or definitions without commentary.
4. Refers exclusively to animal vaccines: discussion is about pets or livestock, not human vaccination.

Output Format: Strictly follow the guidelines above and classify the tweet if it is Favor, Against, or Neutral towards the target "vaccines". Reply with the classification

```
    label only, without any additional text or explanation.

Tweet: {tweet}
Stance: [Favor/Against/Neutral]"""}
]
```

where, in this prompt:

- The `system` message instructs the model to act as a stance detection model with a focus on accurately identifying the author's stance towards vaccines.
- The `user` message contains the specific query for stance detection. It requests a classification of the post into one of three categories: `favor`, `against`, or `neutral`. The model is prompted to provide one of these responses based on the content of the post.

B.2 COVID-19 vaccine timeline

- **January 2020**: The COVID-19 pandemic starts with the identification of a new coronavirus in Wuhan, China.
- **March 2020**: Researchers begin working on developing vaccines for the virus.
- **April 2020**: Clinical trials for several vaccine candidates begin.
- **December 2020**: Emergency use authorization is granted for the first COVID-19 vaccine, developed by Pfizer and BioNTech.
- **January 2021**: More vaccines are granted emergency use authorization, including those developed by Moderna and AstraZeneca.
- **February 2021**: COVID-19 vaccination campaigns begin in many countries.

- **March 2021**: The World Health Organization (WHO) begins to roll out its COVAX initiative, which aims to provide equitable access to COVID-19 vaccines for all countries.
- **April 2021**: The WHO announces that the COVID-19 pandemic is a "once-in-a-century health crisis."
- **May 2021**: The WHO reports that over half of the world's population has received at least one dose of a COVID-19 vaccine.
- **June 2021**: More than 20 COVID-19 vaccine candidates are in clinical trials, and many countries have achieved high levels of vaccination coverage.
- **September 2021**: Some studies suggest that booster shots may be necessary to maintain protection against COVID-19, particularly for certain vaccines and in certain populations.
- **October 2021**: Some manufacturers, such as Pfizer and Moderna, begin clinical trials of booster shots for their respective vaccines.
- **November 2021**: The WHO announces that booster shots may be necessary for certain vaccines, and recommends that countries plan for booster campaigns as needed.
- **December 2021**: Some countries begin to roll out booster shot campaigns for certain populations, such as those who received the AstraZeneca vaccine earlier in the year.
- **January 2022**: Booster shot campaigns become more widespread, with many countries providing booster shots to those who have received a COVID-19 vaccine earlier in the year.
- **February 2022**: The WHO announces that booster shots are effective at increasing immunity to COVID-19, and recommends that countries prioritize booster shots for those at the highest risk of severe illness from the virus.

B.3 Related work

Table A1: Summary of Related Works on Vaccines and COVID-19 post dataset

Authors	Dataset	Total posts	Timeline
Sentiment Analysis			
Hu et al. (2021)	COVID-19 English posts	308,755	Mar 1, 2020 - Feb 28, 2021
Praveen et al. (2021)	COVID-19 posts	73,760	Sep 1 - Dec 31, 2020
De Rosis et al. (2021)	COVID-19 Italian posts	774,407	Feb 17 - Mar 22, 2020
Yousefinaghani et al. (2021)	COVID-19 posts	4,552,652	Jan 1 2020 - Jan 31, 2021
Stella et al. (2020)	COVID-19 Italian posts	101,767	Mar 11 - Mar 17, 2020
Alhuzali et al. (2022)	COVID-19 UK posts	516,427	Feb 1, 2020 - Nov 30, 2021
Qorib et al. (2023)	COVID-19 posts	42,796	Sep 26, 2021 - Nov 7, 2021
Sarirete (2023)	COVID-19 posts	230,623	Dec 16, 2020 - Jan 26, 2021
(Saleh et al., 2023)	COVID-19 posts	2,287,344	Feb 1 - Dec 11, 2020
Burwell et al. (2024)	COVID-19 posts	916,686	Dec 1, 2020 - Feb 28, 2021
Emotion Analysis			
De Rosis et al. (2021)	COVID-19 Italian posts	774,407	Feb 17 - Mar 22, 2020
Stella et al. (2020)	COVID-19 Italian posts	101,767	Mar 11 - Mar 17, 2020
Alhuzali et al. (2022)	COVID-19 UK posts	516,427	Feb 1, 2020 - Nov 30, 2021
Topic Modeling			
Hu et al. (2021)	COVID-19 English posts	308,755	Mar 1, 2020 - Feb 28, 2021
Praveen et al. (2021)	COVID-19 posts	73,760	Sep 1 - Dec 31, 2020
Saleh et al. (2023)	COVID-19 posts	2,287,344	Feb 1 - Dec 11, 2020
Stance Detection			
D'Andrea et al. (2019)	Italian Vaccine posts	112,397	Sep 1, 2016 - Jun 30, 2019
(Giovanni et al., 2022)	French, German, Italian COVID-19 posts	70,352,366	Nov 1, 2020 - Nov 15, 2021
Lindelöf et al. (2023)	COVID-19 posts	16,713,238	Mar 1, 2020 - Jul 31, 2021
Mu et al. (2023)	Labeled COVID-19 English posts	3,101	Oct 2020 - May 2022

B.4 Top 15 low warmth and dominance words

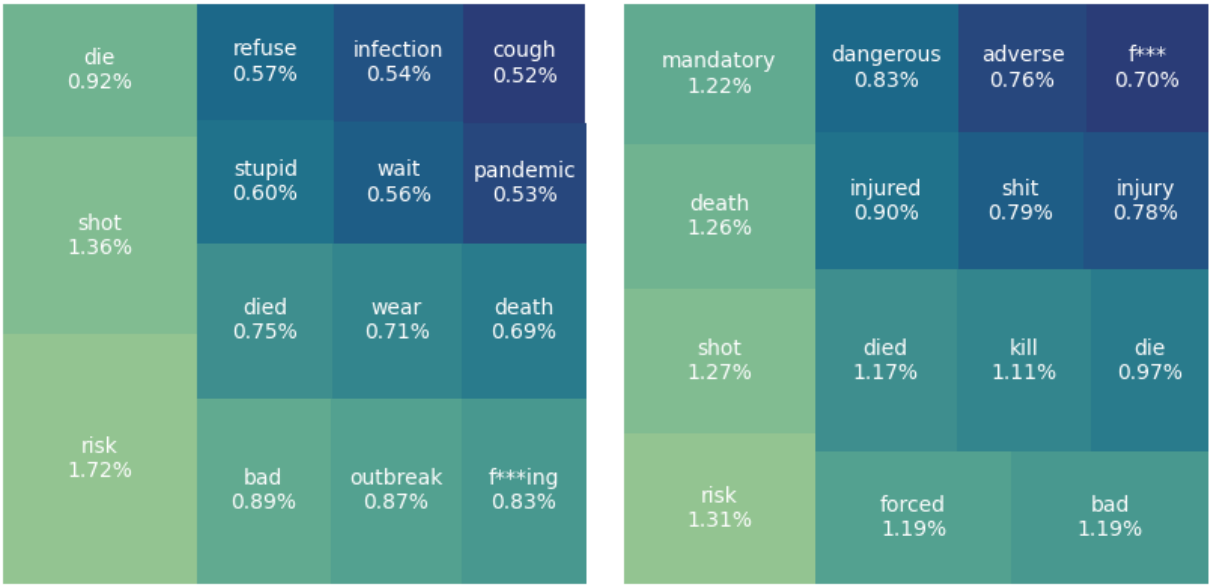


Figure A1: Top 15 low-warmth words for (a) in-favor and (b) against stance towards vaccines.

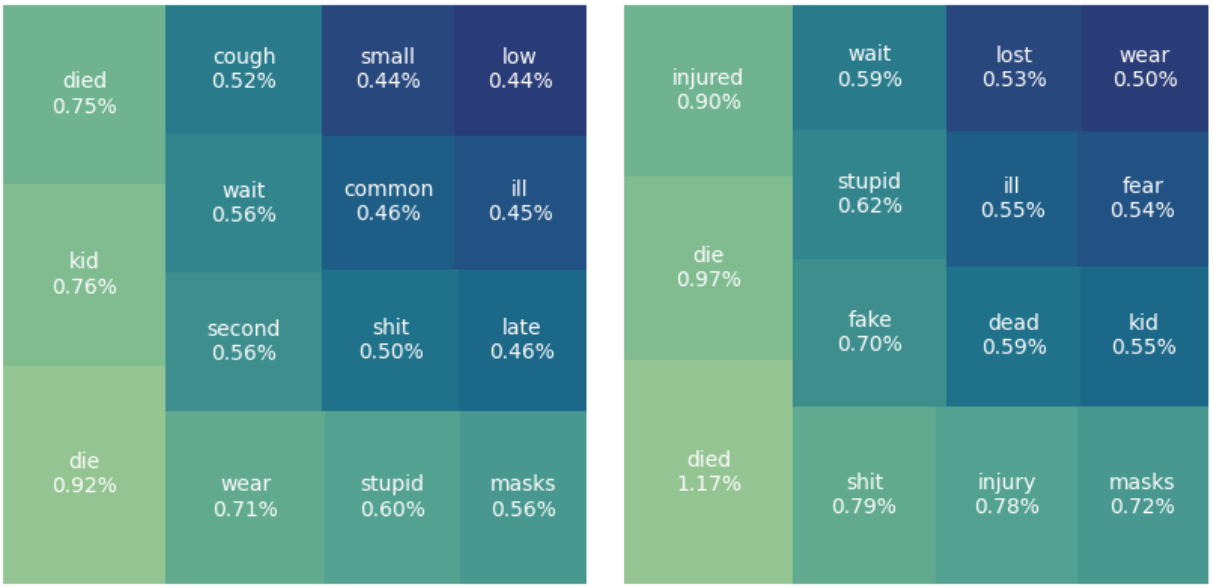


Figure A2: Top 15 low-competence words for (a) in-favor and (b) against stance towards vaccines.

Appendix C

Appendix: Evaluating Large Language Models for stance detection on financial targets from SEC filing reports and earnings call transcripts

C.1 LLM prompt for relevance filtering

```
message = [
        {"role": "system",
```

```
        "content": f"""You are a financial analyst. You are given a sentence
        ↪  that is extracted from the quarterly earnings call transcript of
        ↪  a company. Your task is to classify if the sentence is relevant
        ↪  or not relevant to assess the company's financial outlook
        ↪  towards targets "debt", "eps", or "sales" from profitability
        ↪  perspective. The given sentence should provide sufficient
        ↪  information to make the assessment without needing external
        ↪  context."""},

        {"role": "user",
        "content": f"""
# Task:
Determine if the following sentence from a company's quarterly earnings call
↪  is "relevant" or "not relevant" for assessing the company's financial
↪  outlook toward targets "debt", "eps", or "sales" from a profitability
↪  perspective.

## Classify as 'relevant' if the sentence:

* Directly discusses performance, or outlook for the target.
* States a positive, negative, or neutral stance on the target.
* Contains projections or forward-looking statements about the target.

## Classify as 'not relevant' if the sentence:

* Covers administrative, procedural, or logistical details
* Is a speaker introduction, general background, or technical note
```

```
* Lacks sufficient information to determine the outlook for the target
* Is a legal disclaimer, incomplete sentence, or out of context

**Instructions:**
For the sentence below, determine if the text is **relevant** or **not
 ↪  relevant** for target "{target}". Also, give a short reasoning for your
 ↪  classification.

Sentence: {text}
Relevance: \[relevant / not relevant]
Reason: \[Reasoning for the classification]
"""}]
```

C.2 LLM prompt for stance annotation

```
message = [{"role": "user", "content": f"""
You are a seasoned financial analyst reviewing individual sentences
 ↪  extracted from a company's SEC report. Your task is to determine the
 ↪  company's stance toward a specified financial target **{target}** from a
 ↪  profitability perspective. Classify each sentence as **Positive**,
 ↪  **Negative**, or **Neutral** based on the following criteria:

### Label Definitions:
```

```
* **Positive:** The sentence presents a favorable situation, trend, or
↪  outlook for the target **{target}** that positively impacts or supports
↪  the company's profitability. Examples include reduced debt levels,
↪  rising EPS, or increasing sales.

* **Negative:** The sentence portrays an unfavorable situation, trend, or
↪  outlook for the target **{target}**, negatively impacting or limiting
↪  the company's profitability. Examples include increasing debt burdens,
↪  declining EPS, or falling sales.

* **Neutral:** The sentence does not explicitly or implicitly reflect a
↪  clear stance or impact regarding the target **{target}**, either because
↪  it is not mentioned or because the stance is balanced or unclear.

### Analysis Guidelines:

1. **Focus specifically on mentions or implications related to the financial
↪  target**|avoid general market or macroeconomic comments unless
↪  explicitly tied to the target.
2. **Consider both direct references and contextual implications** (e.g.,
↪  "revenue growth" implicitly refers to sales).
3. **Distinguish clearly between factual reporting and subjective
↪  evaluations**; both types may indicate a stance.
4. **Identify relevant financial cues and indicators**, such as:

   * Performance metrics: growth, stability, decline
```

```
   * Strategic financial decisions: refinancing, restructuring, capital
    ↪  allocation
   * Risk assessments or warnings related to the target
   * Forward-looking statements, future projections, or company guidance
   * Comparative analysis: previous performance or industry benchmarks
5. **Evaluate conflicting or mixed signals** by determining the dominant or
 ↪  most contextually relevant stance.
6. **If the target is neither directly mentioned nor indirectly implied**,
 ↪  classify the stance as **Neutral**.
7. **Even a modest loss per diluted share is still perceived negatively**

{transcript_context}

The sentence for classification is given below:

* Sentence: {text}
* Target: {target}

Provide me the output below:

* Stance: \[Positive / Negative / Neutral]
* Explanation: \[Brief justification of your classification based on the
 ↪  analysis guidelines]
"""}]
```

C.3 Examples of dataset instances

Table A1: Examples of instances in the ECT and SEC datasets for each target and stance class.

ECT example	**SEC example**	**Target**	**Stance**
. . . we're on a path to see adjusted net leverage drop to approximately 3× by year-end, absent any other actions.	*Total debt decreased $6.4 million to $194.4 million at 31 Dec 2020 from $200.8 million a year earlier.*	Debt	Positive
We estimate at least $0.50 of adjusted EPS accretion next year as the business returns toward pre-COVID levels.	*Weighted-average diluted shares outstanding in 2019 declined 2.8%, benefiting earnings per share.*	EPS	Positive
The strong sales mix also led to excellent profitability.	*Selling prices rose 0.6% year-on-year, and lower raw-material costs reduced cost of sales as a percentage of sales.*	Sales	Positive
Adjusted net leverage was 3.7× at year-end, up slightly from Q3 due to input-cost increases impacting EBITDA.	*We cannot assure that operating performance, cash flow, and capital resources will be sufficient to repay our debt.*	Debt	Negative
Combined, these two factors had more than a $0.20 negative impact on adjusted EPS in the quarter.	*Divestitures reduced earnings per diluted share by $0.03 year-on-year for 2020.*	EPS	Negative
We announced a 10% decrease for the segment, driven by two major factors.	*Total sales decreased 14% in Mexico, including a 12% decline in organic sales.*	Sales	Negative
Net debt finished the year at just under $1.2 billion.	*As of 31 Dec 2019, we had approximately $21.6 million of LIBOR-based debt.*	Debt	Neutral
This slide shows the components that impacted our operating margins and EPS versus Q1 last year.	*Discussion of year-on-year changes in operating-income margin and EPS follows: organic growth, productivity, and other.*	EPS	Neutral
Can you sustain that sales momentum?	*Sales grew in home improvement and home care, while consumer health care and stationery and office declined.*	Sales	Neutral

C.4 Chain-of-Thought example

In this section, we present a few chain-of-thought (CoT) reasoning steps generated by the ChatGPT-o3-pro model on a few-shot examples selected from the training set. With the CoT prompt, we prompt LLM to think step-by-step and show the reasoning steps in *Chain-of-Thought Reasoning* section in the output. The following examples show the CoT outputs from the ChatGPT-o3-pro model.

a) Dataset: SEC dataset

Target: earnings per share (EPS)

Instance: On an adjusted basis, net income attributable to 3M was $5.088 billion, or $8.74 per diluted share for 2020 compared to $5.193 billion, or $8.88 per diluted share in December 31, 2019, a decrease of 1.5 percent on a per diluted share basis.

Models output:

1) Chain-of-Thought Reasoning

   - The sentence explicitly discusses "$8.74 per diluted share for 2020" versus "$8.88 per diluted share in 2019," and says this is "a decrease of 1.5 percent."
   - Per Guideline 4 (changes over time), a decline in EPS signals a deterioration in profitability.
   - No offsetting positive cues are given, so the dominant context is the EPS decline.
   - Therefore the stance toward EPS is Negative.
   - The surrounding report data merely corroborate these figures; the sentence itself is sufficient, so the broader context is not necessary for classification.

2) Stance: Negative

3) Brief Justification: The sentence reports adjusted diluted EPS fell from $8.88 to $8.74, a 1.5% decrease, indicating weaker profitability.

b) Dataset: ECT dataset

Target: debt

Instance: all of our debt is prepayable without penalties or fees, and we're using excess cash flow to release borrowings against our revolving credit facility.

1) Chain-of-Thought Reasoning

- The sentence directly addresses debt management.
- "Debt is prepayable without penalties or fees" ⇒ favorable terms; company can retire debt cheaply.
- "Using excess cash flow to release borrowings" ⇒ actively paying down debt, lowering leverage and future interest expense.
- Lower interest expense and greater financial flexibility benefit profitability → Positive stance.

2) Stance: Positive

3) Brief Justification: The sentence shows the company can repay debt inexpensively and is already reducing borrowings with excess cash, a clearly favorable situation for profitability.

C.5 Error analysis

In this section, we do error analysis on the predictions from the LLMs (Llama3:70B). We first look at a few instances where the zero-shot setup makes an incorrect prediction, whereas the few-shot setup gets it correctly. We then look into the cases where the models fail without an explicit CoT prompt but succeed when CoT reasoning is included. Similarly, we look into situations where random examples fail, but similar examples help the model make a correct prediction.

C.5.1 Zero-shot vs. few-shot

1. Dataset: ECT Target: Debt

   Instance: so with ebitda softer than we had anticipated this year reducing our leverage remains a top priority.

   Zero-shot prediction: Positive

   Few-shot ($k = 1$ prediction): Negative

   Ground Truth (ChatGPT-o3-pro model): Negative

   Reasoning from ChatGPT-o3-pro model:

   1) Chain-of-Thought Reasoning

      - "EBITDA softer than we had anticipated" implies earnings have fallen, worsening the firm's leverage ratio.
      - Management says "reducing our leverage remains a top priority," signalling that current debt levels are viewed as too high and potentially harmful to profitability.
      - Emphasis on debt reduction reflects concern, not satisfaction, with the present debt situation. According to the guidelines, such concern over elevated debt is a Negative stance.

   2) Stance: Negative

3) Brief Justification: The sentence states that weaker EBITDA has made lowering leverage a "top priority," indicating current debt levels are a problem for profitability.

C.5.2 CoT vs. non-CoT

1. Dataset: SEC dataset

   Target: sales

   Instance: Research, Development and Related Expenses: R&D, measured as a percent of sales, decreased in 2022 when compared to the same period last year.

   Non-CoT prediction: Positive CoT prediction: Neutral

   Ground-truth (ChatGPT-o3-pro model): Neutral

   Reasoning from ChatGPT-o3-pro model:

   1) Chain-of-Thought Reasoning

      - The only reference to "sales" is as the denominator in the R&D-to-sales ratio.
      - The statement describes a change in R&D spending relative to sales but provides no information about whether sales themselves rose, fell, or remained flat.
      - Because it neither signals a favorable nor an unfavorable development specifically for sales, it fits guideline #6: sales is only tangentially referenced → Neutral.

   2) Stance: Neutral

   3) Brief Justification: The sentence merely notes that R&D expense as a percentage of sales declined, giving no clear indication of sales performance itself.

C.5.3 Random vs. similar example usage

1. Dataset: ECT

Target: earnings per share (EPS)

Instance: we posted organic growth of 0.4% versus our expectation of 1% to 3%, along with adjusted margins of 19% and adjusted earnings of $2.28 per share. Random example prediction: Positive Similar example prediction: Negative Ground-truth (ChatGPT-o3-pro model): Negative

Reasoning from ChatGPT-o3-pro model:

1) Chain-of-Thought Reasoning
    - The sentence compares actual results with management's expectations: organic growth came in at only 0.4% versus the 1%–3% that had been expected.
    - Falling short of internal expectations signals weaker performance and pressure on profitability.
    - Although it lists adjusted EPS of $2.28, there is no positive comparison or improvement stated; the shortfall in growth is the dominant cue.
    - According to Guideline 5, when cues conflict, use the dominant context—in this case, the miss versus expectations—so the stance toward EPS is negative.
2) Stance: Negative
3) Brief Justification: The company delivered just 0.4% organic growth versus its 1%–3% target, implying weaker-than-expected earnings power and a negative outlook for EPS.

C.6 All experiment results

Accuracy (%) of the evaluated models on ECT and SEC datasets under all transcript settings, *with* and *without* chain-of-thought (CoT) reasoning along with *random few-shot sampling* and *most similar few-shot sampling*. The best result in every row is highlighted for each dataset block.

Target	k	ECT Data				SEC Data			
		GPT-4.1-Mini	Gemma3:27B	Llama3.3:70B	Mistral:24B	GPT-4.1-Mini	Gemma3:24B	Llama3.3:70B	Mistral:24B
Random Few-Shot Sampling									
Without Transcript – Without CoT									
debt	0	85.57 ± 1.79	**87.29 ± 0.6**	79.73 ± 3.31	78.01 ± 6.86	**85.91 ± 0.6**	75.6 ± 0.6	78.35 ± 1.79	50.52 ± 3.57
	1	83.51 ± 1.03	**86.6 ± 0.0**	82.13 ± 2.15	76.98 ± 1.19	**82.47 ± 1.79**	75.26 ± 1.03	73.2 ± 1.03	54.3 ± 2.98
	5	83.85 ± 0.6	**84.19 ± 2.15**	**84.19 ± 2.38**	74.23 ± 1.79	**81.79 ± 0.6**	72.51 ± 0.6	74.23 ± 1.03	57.39 ± 2.15
	10	81.44 ± 0.0	**86.25 ± 0.6**	81.79 ± 2.38	71.82 ± 2.38	**80.07 ± 0.6**	73.54 ± 1.19	71.48 ± 2.15	57.04 ± 2.38
eps	0	**88.5 ± 0.68**	83.04 ± 0.0	82.85 ± 1.47	78.36 ± 0.58	**80.18 ± 1.56**	68.47 ± 3.12	72.07 ± 4.13	60.36 ± 1.56
	1	**85.77 ± 0.89**	82.07 ± 0.34	82.46 ± 1.55	75.63 ± 2.36	75.68 ± 0.0	65.77 ± 1.56	**77.48 ± 3.12**	55.86 ± 5.63
	5	**87.13 ± 0.0**	81.29 ± 0.58	82.65 ± 0.89	73.29 ± 0.34	**79.28 ± 1.56**	68.47 ± 1.56	76.58 ± 1.56	53.15 ± 3.12
	10	**85.38 ± 1.01**	84.21 ± 0.0	81.87 ± 1.75	74.07 ± 5.11	**79.28 ± 4.13**	67.57 ± 0.0	76.58 ± 1.56	58.56 ± 4.13
sales	0	**92.2 ± 0.34**	88.36 ± 0.2	85.88 ± 1.99	80.9 ± 2.15	91.62 ± 0.34	**92.2 ± 0.34**	89.08 ± 1.88	78.36 ± 1.55
	1	**91.19 ± 0.34**	87.91 ± 0.71	87.23 ± 1.74	79.21 ± 1.67	**91.81 ± 0.58**	89.86 ± 0.34	86.74 ± 4.11	76.41 ± 1.79
	5	**90.73 ± 0.2**	88.93 ± 0.2	89.15 ± 0.34	80.34 ± 1.22	**92.2 ± 1.35**	87.33 ± 0.89	88.3 ± 1.75	74.27 ± 2.34
	10	**91.53 ± 0.59**	89.27 ± 0.2	87.8 ± 2.69	82.82 ± 2.82	**90.25 ± 0.34**	88.69 ± 0.68	87.72 ± 2.55	77.19 ± 2.55
Without Transcript – With CoT									
debt	0	80.41 ± 3.72	68.38 ± 1.57	**83.16 ± 2.59**	77.32 ± 2.06	**82.47 ± 3.72**	69.42 ± 2.98	78.35 ± 1.79	72.85 ± 7.81
	1	81.1 ± 2.38	83.16 ± 3.15	**83.51 ± 2.06**	78.01 ± 5.68	**82.47 ± 2.73**	79.38 ± 2.06	78.69 ± 1.57	72.51 ± 4.87
	5	**86.94 ± 1.57**	85.57 ± 1.79	83.16 ± 1.57	78.01 ± 5.29	81.79 ± 1.57	82.82 ± 4.17	**83.16 ± 3.62**	65.64 ± 3.9
	10	**85.91 ± 2.59**	85.22 ± 3.31	83.51 ± 0.0	73.54 ± 6.55	**84.54 ± 2.06**	83.85 ± 0.6	82.82 ± 2.38	75.26 ± 3.57
eps	0	**81.68 ± 2.94**	51.27 ± 0.68	80.51 ± 1.88	79.92 ± 2.88	**87.39 ± 4.13**	45.05 ± 6.24	72.97 ± 2.7	71.17 ± 3.12
	1	**86.94 ± 1.47**	85.96 ± 0.58	83.63 ± 1.17	82.07 ± 2.64	**91.89 ± 5.41**	74.77 ± 1.56	79.28 ± 1.56	63.96 ± 7.8
	5	**88.11 ± 1.22**	87.13 ± 1.01	83.04 ± 2.03	77.97 ± 1.22	**88.29 ± 3.12**	69.37 ± 7.8	78.38 ± 5.41	74.77 ± 3.12
	10	**85.58 ± 2.21**	**85.58 ± 2.05**	84.21 ± 0.58	80.51 ± 1.47	**85.59 ± 4.13**	70.27 ± 0.0	72.07 ± 1.56	72.97 ± 2.7
sales	0	78.98 ± 5.33	62.6 ± 3.41	**87.34 ± 1.28**	83.73 ± 1.55	84.6 ± 1.79	63.94 ± 3.0	**91.03 ± 0.68**	84.21 ± 3.56
	1	**92.54 ± 1.17**	90.96 ± 0.71	90.62 ± 1.74	85.42 ± 2.37	**90.25 ± 2.43**	**90.25 ± 0.34**	**90.25 ± 0.89**	84.02 ± 1.47
	5	**92.54 ± 1.36**	90.06 ± 0.85	90.73 ± 0.52	82.82 ± 1.67	88.3 ± 2.11	**89.47 ± 1.55**	88.5 ± 1.22	80.9 ± 1.88
	10	**92.66 ± 0.71**	91.41 ± 0.71	89.72 ± 0.52	83.84 ± 1.41	**90.25 ± 0.68**	88.3 ± 0.58	88.89 ± 1.55	82.85 ± 1.47

Continued on next page

Continued from previous page

Target	k	ECT Data				SEC Data			
		GPT-4.1-Mini	Llama3.3:70B	Gemma3:27B	Mistral:24B	GPT-4.1-Mini	Llama3.3:70B	Gemma3:24B	Mistral:24B
Full Transcript – Without CoT									
debt	0	**92.1 ± 0.6**	84.19 ± 1.57	87.63 ± 2.73	40.89 ± 2.38	**82.82 ± 0.6**	57.39 ± 1.19	55.33 ± 0.6	25.09 ± 1.57
	1	**89.35 ± 0.6**	79.73 ± 1.57	86.94 ± 3.62	41.92 ± 3.15	**79.73 ± 1.57**	62.54 ± 1.57	59.79 ± 2.06	28.18 ± 2.15
	5	**89.69 ± 1.79**	83.85 ± 0.6	84.54 ± 1.79	45.02 ± 5.68	**79.38 ± 1.03**	64.6 ± 1.57	56.36 ± 2.15	33.33 ± 0.6
	10	**91.41 ± 0.6**	83.51 ± 0.0	84.54 ± 3.09	46.05 ± 1.19	**80.41 ± 0.0**	65.64 ± 1.57	55.33 ± 0.6	36.08 ± 2.06
eps	0	**87.52 ± 0.68**	86.74 ± 1.22	86.74 ± 2.21	63.74 ± 1.55	**81.08 ± 0.0**	65.77 ± 1.56	65.77 ± 1.56	44.14 ± 8.26
	1	**89.08 ± 0.68**	87.91 ± 1.35	86.16 ± 2.43	65.5 ± 1.55	**81.08 ± 0.0**	70.27 ± 0.0	64.86 ± 0.0	56.76 ± 7.15
	5	**89.47 ± 0.58**	84.8 ± 1.55	85.96 ± 2.11	64.52 ± 0.68	**81.08 ± 0.0**	69.37 ± 1.56	63.06 ± 4.13	47.75 ± 3.12
	10	**90.25 ± 0.34**	85.19 ± 0.89	83.43 ± 2.05	67.06 ± 2.05	**79.28 ± 1.56**	66.67 ± 4.13	63.96 ± 3.12	57.66 ± 4.13
sales	0	**92.77 ± 0.2**	89.38 ± 0.52	89.49 ± 1.76	64.07 ± 2.35	**91.62 ± 1.22**	85.38 ± 1.75	87.52 ± 0.89	57.31 ± 0.58
	1	**93.9 ± 0.59**	89.72 ± 0.71	88.47 ± 1.79	66.78 ± 1.89	**92.59 ± 0.34**	88.69 ± 0.89	87.91 ± 0.89	71.15 ± 4.15
	5	**94.35 ± 0.52**	88.47 ± 0.34	88.7 ± 1.87	70.4 ± 2.21	**90.25 ± 0.89**	86.55 ± 2.34	86.16 ± 1.88	64.91 ± 1.55
	10	**92.66 ± 0.85**	88.59 ± 1.04	87.46 ± 0.59	67.12 ± 0.9	**89.86 ± 0.34**	84.6 ± 3.71	86.16 ± 2.76	66.08 ± 3.26
Full Transcript – With CoT									
debt	0	**90.03 ± 1.57**	80.07 ± 2.59	84.88 ± 2.15	50.86 ± 2.15	75.6 ± 3.15	67.35 ± 1.57	**78.01 ± 1.57**	34.71 ± 4.65
	1	**91.07 ± 2.38**	85.91 ± 1.57	84.54 ± 1.79	81.1 ± 3.9	76.98 ± 4.65	**78.01 ± 1.57**	**78.01 ± 4.17**	68.38 ± 0.6
	5	**92.1 ± 2.59**	84.54 ± 3.72	87.29 ± 3.15	78.01 ± 5.19	**81.1 ± 1.57**	**81.1 ± 1.57**	78.69 ± 2.15	69.42 ± 2.38
	10	**92.44 ± 0.6**	85.91 ± 1.19	87.63 ± 3.72	75.6 ± 1.19	**81.44 ± 2.73**	80.76 ± 2.15	79.73 ± 1.57	74.23 ± 2.73
eps	0	**85.77 ± 1.79**	81.09 ± 0.89	85.19 ± 1.79	56.53 ± 2.36	**78.38 ± 7.15**	59.46 ± 5.41	68.47 ± 3.12	39.64 ± 1.56
	1	**91.81 ± 1.75**	88.11 ± 0.89	85.96 ± 2.68	79.53 ± 1.17	**90.99 ± 1.56**	73.87 ± 3.12	72.07 ± 4.13	68.47 ± 4.13
	5	**89.28 ± 1.69**	86.35 ± 0.34	85.96 ± 2.34	79.14 ± 1.47	**87.39 ± 6.8**	77.48 ± 3.12	75.68 ± 2.7	66.67 ± 1.56
	10	**88.69 ± 0.89**	86.94 ± 0.68	85.19 ± 1.79	80.7 ± 1.55	**88.29 ± 4.13**	72.97 ± 2.7	77.48 ± 3.12	68.47 ± 1.56
sales	0	90.17 ± 1.79	81.58 ± 2.76	**90.62 ± 1.28**	67.12 ± 1.22	**89.47 ± 1.17**	67.84 ± 2.55	88.89 ± 1.55	59.45 ± 4.11
	1	**94.24 ± 1.48**	91.19 ± 1.02	88.93 ± 1.04	87.12 ± 1.17	**90.25 ± 0.89**	89.08 ± 1.88	87.13 ± 2.11	84.21 ± 2.11
	5	**94.12 ± 1.37**	90.51 ± 0.9	89.49 ± 1.17	85.08 ± 1.36	**91.03 ± 1.47**	89.28 ± 1.79	84.6 ± 1.47	83.24 ± 0.68
	10	**94.24 ± 1.22**	91.19 ± 0.34	89.83 ± 2.12	84.41 ± 1.02	**90.06 ± 1.75**	**90.06 ± 0.0**	85.96 ± 1.75	82.85 ± 1.47
Summarized Transcript – Without CoT									
debt	0	**90.38 ± 0.6**	88.32 ± 0.6	84.19 ± 4.76	76.98 ± 0.6	**82.47 ± 1.03**	71.82 ± 0.6	75.6 ± 2.15	21.31 ± 2.15
	1	**90.03 ± 0.6**	87.29 ± 1.57	87.29 ± 0.6	69.42 ± 2.98	**84.54 ± 1.03**	70.45 ± 2.15	71.82 ± 1.57	21.99 ± 2.15
	5	**88.66 ± 1.03**	82.13 ± 0.6	85.22 ± 3.31	61.51 ± 3.62	**82.13 ± 0.6**	67.7 ± 1.57	79.04 ± 2.15	30.58 ± 0.6
	10	**88.66 ± 1.03**	84.88 ± 1.19	85.22 ± 4.17	55.33 ± 3.15	**82.13 ± 0.6**	69.42 ± 0.6	80.76 ± 2.59	36.43 ± 2.15
eps	0	**88.69 ± 0.34**	84.02 ± 0.89	85.38 ± 1.17	71.93 ± 3.09	**81.08 ± 2.7**	69.37 ± 4.13	67.57 ± 4.68	40.54 ± 5.41

Continued on next page

Continued from previous page

Target	k	ECT Data				SEC Data			
		GPT-4.1-Mini	Llama3.3:70B	Gemma3:27B	Mistral:24B	GPT-4.1-Mini	Llama3.3:70B	Gemma3:24B	Mistral:24B
	1	**87.33 ± 0.68**	84.99 ± 0.34	84.6 ± 2.05	68.42 ± 3.26	**81.98 ± 1.56**	70.27 ± 2.7	74.77 ± 1.56	45.95 ± 2.7
	5	**87.52 ± 0.34**	85.77 ± 0.34	84.8 ± 2.55	65.5 ± 1.75	**84.68 ± 1.56**	63.06 ± 4.13	75.68 ± 2.7	51.35 ± 2.7
	10	**87.72 ± 0.0**	85.38 ± 0.0	84.8 ± 1.17	66.86 ± 1.47	**81.98 ± 1.56**	69.37 ± 6.24	72.07 ± 1.56	53.15 ± 3.12
sales	0	**93.22 ± 0.68**	89.83 ± 0.68	88.59 ± 2.54	74.01 ± 2.07	**92.4 ± 0.58**	87.91 ± 1.88	89.86 ± 2.36	45.81 ± 3.38
	1	**93.11 ± 0.52**	90.62 ± 0.52	87.8 ± 2.94	71.3 ± 1.57	**91.42 ± 1.47**	90.64 ± 1.17	**91.42 ± 2.88**	55.17 ± 0.89
	5	**93.9 ± 0.0**	90.4 ± 0.39	88.59 ± 2.59	68.7 ± 0.52	**91.62 ± 0.34**	87.13 ± 1.55	**91.62 ± 2.76**	57.12 ± 2.88
	10	**93.9 ± 0.34**	90.85 ± 0.68	88.81 ± 2.9	67.8 ± 1.89	90.45 ± 0.68	89.28 ± 0.89	**92.4 ± 0.58**	64.13 ± 2.43
Summarized Transcript – With CoT									
debt	0	87.29 ± 2.59	78.69 ± 2.59	**88.32 ± 1.19**	79.73 ± 2.15	77.66 ± 2.15	66.32 ± 1.57	**81.44 ± 4.72**	38.14 ± 4.72
	1	**89.0 ± 2.98**	86.6 ± 1.03	85.57 ± 2.06	79.04 ± 8.27	79.38 ± 2.06	80.41 ± 1.79	**81.1 ± 2.38**	70.45 ± 4.29
	5	87.97 ± 2.59	85.22 ± 2.59	**88.32 ± 1.57**	77.66 ± 1.57	**82.47 ± 1.03**	82.13 ± 1.19	82.13 ± 0.6	68.73 ± 3.31
	10	**86.94 ± 3.62**	84.19 ± 2.15	83.85 ± 0.6	76.29 ± 3.57	**84.19 ± 2.15**	80.41 ± 2.73	81.79 ± 2.38	79.38 ± 1.79
eps	0	**89.86 ± 2.05**	74.66 ± 0.68	84.02 ± 0.34	70.37 ± 3.22	**84.68 ± 7.8**	60.36 ± 1.56	75.68 ± 4.68	41.44 ± 6.24
	1	**91.03 ± 1.69**	87.91 ± 0.68	89.08 ± 1.22	82.46 ± 0.58	**89.19 ± 4.68**	73.87 ± 1.56	76.58 ± 1.56	71.17 ± 4.13
	5	**87.72 ± 1.55**	86.55 ± 1.17	86.55 ± 1.55	79.92 ± 1.88	**89.19 ± 5.41**	74.77 ± 8.26	77.48 ± 1.56	68.47 ± 10.92
	10	**88.3 ± 0.58**	86.94 ± 1.35	83.24 ± 1.35	78.75 ± 2.36	**88.29 ± 3.12**	74.77 ± 3.12	75.68 ± 2.7	69.37 ± 7.8
sales	0	**93.45 ± 0.2**	77.97 ± 1.48	88.47 ± 1.89	78.08 ± 2.4	89.28 ± 0.89	75.83 ± 0.34	**89.86 ± 0.68**	54.58 ± 1.35
	1	**92.99 ± 1.37**	90.28 ± 0.2	91.41 ± 1.19	86.78 ± 1.55	90.64 ± 1.17	89.28 ± 2.05	**90.84 ± 0.34**	84.41 ± 2.21
	5	92.09 ± 0.71	90.62 ± 0.98	**92.32 ± 1.04**	84.18 ± 1.09	**90.84 ± 1.47**	89.67 ± 1.79	89.47 ± 0.58	82.26 ± 2.36
	10	**93.33 ± 0.71**	90.51 ± 0.34	91.41 ± 1.28	84.41 ± 2.9	89.67 ± 0.89	**89.86 ± 1.22**	89.08 ± 1.47	83.24 ± 0.34
Similar Few-Shot Sampling									
Without Transcript – Without CoT									
debt	0	85.57 ± 1.79	**87.29 ± 0.6**	84.19 ± 2.15	78.69 ± 4.65	**84.88 ± 1.19**	75.6 ± 0.6	76.29 ± 1.79	53.26 ± 3.31
	1	83.85 ± 0.6	85.22 ± 0.6	**85.57 ± 2.73**	79.73 ± 4.29	**83.51 ± 1.79**	78.35 ± 1.79	75.95 ± 1.19	58.08 ± 2.59
	5	80.76 ± 0.6	**86.25 ± 0.6**	84.54 ± 2.73	77.32 ± 1.03	**82.82 ± 0.6**	74.91 ± 1.57	79.73 ± 2.15	62.89 ± 2.06
	10	83.16 ± 2.15	**85.22 ± 0.6**	83.16 ± 1.57	80.07 ± 2.59	**84.19 ± 0.6**	78.01 ± 0.6	73.54 ± 0.6	61.51 ± 1.57
eps	0	**88.89 ± 1.01**	82.26 ± 0.34	83.63 ± 1.01	77.0 ± 0.34	**81.98 ± 1.56**	71.17 ± 1.56	75.68 ± 0.0	53.15 ± 5.63
	1	**88.3 ± 1.01**	84.99 ± 0.68	87.13 ± 2.03	77.58 ± 2.21	**79.28 ± 3.12**	68.47 ± 1.56	72.97 ± 0.0	58.56 ± 1.56
	5	**87.33 ± 0.89**	84.8 ± 0.58	87.13 ± 3.83	79.73 ± 0.89	**77.48 ± 1.56**	66.67 ± 1.56	69.37 ± 6.8	60.36 ± 1.56
	10	**87.13 ± 0.58**	86.55 ± 0.0	**87.13 ± 0.58**	78.36 ± 1.01	**80.18 ± 1.56**	67.57 ± 0.0	73.87 ± 4.13	55.86 ± 8.69
sales	0	**92.09 ± 0.52**	88.7 ± 0.2	85.88 ± 1.67	79.1 ± 0.2	**91.81 ± 0.58**	91.62 ± 0.89	90.84 ± 0.34	77.39 ± 1.88

Continued on next page

Continued from previous page

Target	k	ECT Data				SEC Data			
		GPT-4.1-Mini	Llama3.3:70B	Gemma3:27B	Mistral:24B	GPT-4.1-Mini	Llama3.3:70B	Gemma3:24B	Mistral:24B
	1	**93.33 ± 0.2**	88.7 ± 0.52	90.73 ± 1.6	82.15 ± 1.67	**94.93 ± 0.34**	93.96 ± 0.68	89.86 ± 2.94	82.85 ± 2.76
	5	**92.32 ± 0.39**	87.46 ± 0.0	88.59 ± 2.74	83.28 ± 1.41	**91.81 ± 0.58**	**91.81 ± 0.58**	89.08 ± 3.0	80.31 ± 2.05
	10	**92.2 ± 0.34**	88.7 ± 0.52	89.38 ± 0.85	82.82 ± 3.15	**93.37 ± 0.68**	93.18 ± 0.34	90.45 ± 0.34	83.24 ± 0.34
Without Transcript – With CoT									
debt	0	81.79 ± 3.15	72.16 ± 4.12	**85.22 ± 1.19**	78.69 ± 2.59	**82.82 ± 1.57**	63.57 ± 1.19	80.07 ± 2.38	75.6 ± 3.62
	1	**85.57 ± 2.73**	84.54 ± 1.79	84.88 ± 2.15	76.98 ± 1.19	**85.57 ± 2.73**	84.19 ± 2.38	84.54 ± 2.06	74.57 ± 1.19
	5	**84.88 ± 2.59**	81.44 ± 2.06	80.07 ± 1.19	78.35 ± 1.03	**89.35 ± 2.15**	85.91 ± 2.15	87.29 ± 2.15	70.1 ± 4.49
	10	**87.97 ± 2.15**	85.57 ± 1.79	82.13 ± 1.57	76.63 ± 2.15	**88.32 ± 2.38**	83.85 ± 1.57	82.47 ± 3.09	79.38 ± 5.36
eps	0	**80.51 ± 3.0**	53.02 ± 2.64	80.12 ± 2.03	78.36 ± 1.75	**86.49 ± 2.7**	37.84 ± 5.41	80.18 ± 4.13	63.06 ± 6.24
	1	**87.52 ± 0.34**	87.33 ± 1.69	87.13 ± 0.58	79.53 ± 1.75	**92.79 ± 3.12**	82.88 ± 4.13	82.88 ± 3.12	74.77 ± 5.63
	5	85.58 ± 0.89	84.6 ± 1.35	**86.74 ± 0.68**	79.73 ± 2.05	**88.29 ± 6.8**	78.38 ± 2.7	83.78 ± 2.7	68.47 ± 11.25
	10	84.6 ± 0.89	**86.35 ± 0.89**	85.96 ± 2.92	79.92 ± 2.43	**90.09 ± 4.13**	72.07 ± 3.12	77.48 ± 4.13	60.36 ± 1.56
sales	0	77.85 ± 5.45	59.77 ± 2.31	**88.59 ± 0.52**	83.05 ± 1.02	86.74 ± 2.21	63.35 ± 1.79	**89.28 ± 0.89**	85.19 ± 0.89
	1	**92.66 ± 0.85**	88.93 ± 0.2	90.06 ± 1.28	87.68 ± 1.87	93.37 ± 1.22	**95.71 ± 0.34**	94.54 ± 0.34	90.45 ± 1.22
	5	**92.09 ± 0.85**	90.4 ± 1.67	89.94 ± 0.39	85.65 ± 1.37	92.79 ± 2.05	**94.15 ± 2.11**	92.59 ± 0.89	85.96 ± 2.03
	10	**92.43 ± 0.39**	90.73 ± 1.09	89.72 ± 0.85	86.89 ± 0.52	91.81 ± 1.01	**95.13 ± 0.89**	92.01 ± 2.94	85.77 ± 1.22
Full Transcript – Without CoT									
debt	0	**91.41 ± 1.57**	84.54 ± 1.79	88.32 ± 1.57	39.52 ± 2.98	**82.47 ± 1.03**	58.42 ± 2.59	51.89 ± 2.15	24.4 ± 1.57
	1	**89.35 ± 1.19**	80.41 ± 1.79	84.54 ± 2.73	58.76 ± 1.03	**79.73 ± 1.57**	68.04 ± 2.73	61.86 ± 2.06	45.36 ± 2.06
	5	**89.0 ± 1.19**	80.41 ± 1.03	82.47 ± 3.09	64.95 ± 2.06	**79.73 ± 0.6**	67.7 ± 2.38	65.29 ± 1.19	51.89 ± 2.38
	10	**85.91 ± 1.57**	81.44 ± 1.03	82.47 ± 3.72	61.17 ± 11.5	**80.76 ± 1.57**	67.01 ± 1.03	62.2 ± 2.15	52.58 ± 1.03
eps	0	**87.52 ± 0.68**	86.35 ± 0.89	85.77 ± 3.22	62.18 ± 1.88	**81.08 ± 0.0**	66.67 ± 1.56	63.06 ± 1.56	42.34 ± 4.13
	1	**91.81 ± 0.58**	87.33 ± 0.34	85.77 ± 2.05	67.64 ± 2.36	**84.68 ± 1.56**	71.17 ± 1.56	67.57 ± 0.0	48.65 ± 2.7
	5	**91.03 ± 0.34**	85.96 ± 0.0	84.8 ± 2.55	70.57 ± 3.22	**81.08 ± 2.7**	72.97 ± 2.7	67.57 ± 2.7	54.95 ± 4.13
	10	**90.84 ± 0.89**	84.02 ± 1.22	84.8 ± 1.55	69.2 ± 1.47	**81.98 ± 3.12**	73.87 ± 1.56	67.57 ± 0.0	55.86 ± 1.56
sales	0	**92.99 ± 0.52**	89.38 ± 1.09	88.47 ± 2.71	62.37 ± 1.79	**91.42 ± 0.89**	86.55 ± 1.55	88.11 ± 2.21	60.43 ± 3.0
	1	**94.01 ± 0.2**	88.02 ± 0.52	88.93 ± 2.21	65.88 ± 4.58	**93.76 ± 0.34**	86.94 ± 1.79	90.25 ± 1.79	77.19 ± 1.75
	5	**93.79 ± 0.2**	89.04 ± 0.85	90.28 ± 1.37	72.09 ± 1.19	**91.03 ± 1.47**	87.13 ± 2.11	88.3 ± 2.11	76.22 ± 2.88
	10	**93.77 ± 0.22**	89.49 ± 0.59	89.94 ± 1.41	69.04 ± 0.85	**89.86 ± 0.34**	87.72 ± 2.55	86.94 ± 2.21	74.07 ± 1.69
Full Transcript – With CoT									
debt	0	**87.97 ± 4.87**	74.91 ± 4.17	84.19 ± 1.57	50.86 ± 5.86	**75.6 ± 0.6**	68.04 ± 3.57	74.91 ± 2.15	36.77 ± 5.68
	1	**91.75 ± 1.03**	81.79 ± 2.15	84.19 ± 1.57	79.73 ± 3.62	**86.94 ± 2.15**	83.51 ± 1.79	84.54 ± 2.06	76.98 ± 2.59

Continued on next page

Continued from previous page

Target	k	ECT Data				SEC Data			
		GPT-4.1-Mini	Llama3.3:70B	Gemma3:27B	Mistral:24B	GPT-4.1-Mini	Llama3.3:70B	Gemma3:24B	Mistral:24B
	5	**90.38 ± 0.6**	82.82 ± 1.19	84.88 ± 2.15	80.07 ± 2.15	**89.35 ± 1.57**	83.51 ± 1.03	83.85 ± 2.59	76.98 ± 3.31
	10	**91.07 ± 1.57**	85.22 ± 1.57	85.57 ± 3.09	80.07 ± 3.62	**86.94 ± 1.57**	83.85 ± 0.6	84.19 ± 5.68	80.07 ± 4.17
eps	0	**86.55 ± 1.55**	84.6 ± 1.35	84.6 ± 0.89	59.65 ± 2.55	**79.28 ± 4.13**	63.96 ± 4.13	72.97 ± 2.7	46.85 ± 6.24
	1	**90.84 ± 0.68**	88.3 ± 0.58	85.38 ± 0.58	80.7 ± 1.55	**91.89 ± 2.7**	81.98 ± 1.56	75.68 ± 2.7	73.87 ± 1.56
	5	**89.86 ± 1.35**	88.5 ± 0.89	84.8 ± 1.55	84.21 ± 1.17	**83.78 ± 2.7**	81.08 ± 2.7	77.48 ± 1.56	66.67 ± 8.26
	10	**90.45 ± 0.89**	87.72 ± 1.17	84.6 ± 0.89	80.51 ± 3.57	**85.59 ± 1.56**	75.68 ± 2.7	75.68 ± 2.7	64.86 ± 2.7
sales	0	89.27 ± 1.99	81.58 ± 1.19	**91.86 ± 2.12**	64.52 ± 2.92	**90.25 ± 3.0**	66.47 ± 2.64	89.08 ± 0.68	60.43 ± 4.23
	1	**95.48 ± 0.78**	90.96 ± 1.04	90.4 ± 0.52	87.68 ± 2.54	94.15 ± 2.34	**95.32 ± 0.58**	90.25 ± 1.47	91.23 ± 0.58
	5	**93.9 ± 0.34**	90.4 ± 0.52	90.85 ± 1.36	85.88 ± 1.28	93.57 ± 1.17	**94.35 ± 0.68**	88.69 ± 1.22	86.55 ± 1.55
	10	**93.33 ± 0.85**	91.19 ± 0.68	90.51 ± 0.68	87.23 ± 1.28	**93.76 ± 0.68**	93.18 ± 0.34	87.72 ± 1.55	84.6 ± 1.22
Summarized Transcript – Without CoT									
debt	0	**89.35 ± 0.6**	88.32 ± 0.6	84.88 ± 3.31	74.23 ± 4.72	**82.82 ± 1.57**	71.82 ± 0.6	76.29 ± 2.06	20.27 ± 0.6
	1	**89.0 ± 1.57**	86.6 ± 1.03	85.91 ± 3.62	75.26 ± 1.79	**82.47 ± 1.03**	73.54 ± 0.6	78.35 ± 1.79	44.33 ± 1.03
	5	**87.29 ± 0.6**	84.54 ± 1.03	83.51 ± 4.49	69.76 ± 2.15	**82.82 ± 0.6**	71.48 ± 0.6	81.79 ± 1.57	47.77 ± 2.38
	10	**89.0 ± 0.6**	83.85 ± 1.57	85.22 ± 2.15	67.01 ± 1.03	**82.47 ± 0.0**	69.76 ± 3.31	76.98 ± 3.31	49.83 ± 2.59
eps	0	**88.89 ± 0.58**	84.6 ± 1.47	84.8 ± 2.11	70.57 ± 1.79	**79.28 ± 3.12**	71.17 ± 1.56	68.47 ± 3.12	46.85 ± 6.8
	1	**88.89 ± 1.17**	85.38 ± 0.0	84.02 ± 3.22	70.96 ± 2.64	**85.59 ± 4.13**	70.27 ± 2.7	68.47 ± 3.12	50.45 ± 3.12
	5	**89.28 ± 0.89**	87.33 ± 0.34	84.6 ± 2.43	70.57 ± 2.76	**83.78 ± 2.7**	72.97 ± 4.68	72.07 ± 3.12	60.36 ± 1.56
	10	**89.08 ± 0.34**	87.13 ± 0.0	85.19 ± 1.35	73.29 ± 2.21	**83.78 ± 0.0**	74.77 ± 4.13	69.37 ± 1.56	54.95 ± 6.24
sales	0	**92.88 ± 0.34**	90.17 ± 0.68	89.94 ± 2.07	71.86 ± 0.68	**91.81 ± 0.58**	88.11 ± 1.22	91.23 ± 1.55	44.44 ± 1.55
	1	**93.11 ± 0.2**	89.94 ± 0.39	88.47 ± 2.03	76.05 ± 0.52	**93.76 ± 0.68**	91.23 ± 0.58	89.47 ± 2.92	65.69 ± 3.8
	5	**93.45 ± 0.39**	88.93 ± 0.2	89.72 ± 1.6	73.33 ± 1.28	92.59 ± 0.34	89.67 ± 0.89	**92.79 ± 1.22**	69.79 ± 1.22
	10	**92.77 ± 0.85**	89.72 ± 0.85	88.02 ± 1.41	66.33 ± 3.62	**92.01 ± 0.34**	90.06 ± 1.55	89.08 ± 0.34	72.32 ± 2.64
Summarized Transcript – With CoT									
debt	0	**88.32 ± 2.59**	80.07 ± 2.98	86.25 ± 3.62	75.95 ± 1.57	79.73 ± 1.19	69.76 ± 3.15	**80.76 ± 0.6**	35.74 ± 0.6
	1	**88.32 ± 2.15**	81.44 ± 0.0	85.22 ± 1.57	75.26 ± 3.09	**86.25 ± 1.19**	84.54 ± 1.79	84.88 ± 1.19	80.07 ± 4.87
	5	**90.03 ± 2.15**	82.13 ± 0.6	84.54 ± 2.06	80.07 ± 2.98	**89.35 ± 1.57**	85.57 ± 1.79	85.57 ± 1.03	72.85 ± 4.29
	10	**89.69 ± 1.79**	83.51 ± 1.03	85.91 ± 2.15	78.69 ± 1.57	**89.69 ± 1.03**	82.13 ± 2.59	82.82 ± 3.31	80.41 ± 2.06
eps	0	**87.91 ± 1.22**	74.07 ± 0.89	86.35 ± 0.89	73.68 ± 3.26	**88.29 ± 1.56**	59.46 ± 0.0	70.27 ± 2.7	45.95 ± 14.04
	1	**89.67 ± 0.34**	89.08 ± 2.76	86.94 ± 0.89	79.14 ± 0.68	**92.79 ± 3.12**	81.98 ± 1.56	81.08 ± 2.7	72.97 ± 2.7
	5	**90.25 ± 1.69**	88.11 ± 1.22	85.58 ± 2.21	82.07 ± 0.34	**91.89 ± 0.0**	81.98 ± 3.12	81.08 ± 2.7	68.47 ± 4.13
	10	**89.47 ± 0.58**	85.96 ± 0.58	85.58 ± 1.79	79.73 ± 1.22	**87.39 ± 1.56**	80.18 ± 1.56	80.18 ± 5.63	68.47 ± 3.12

Continued on next page

Continued from previous page

Target	k	ECT Data				SEC Data			
		GPT-4.1-Mini	Llama3.3:70B	Gemma3:27B	Mistral:24B	GPT-4.1-Mini	Llama3.3:70B	Gemma3:24B	Mistral:24B
sales	0	**92.88 ± 1.22**	78.42 ± 0.78	88.47 ± 1.22	78.87 ± 1.96	90.25 ± 0.89	74.66 ± 0.34	**90.64 ± 1.01**	50.1 ± 1.35
	1	**93.22 ± 0.68**	91.07 ± 0.71	91.86 ± 0.59	86.44 ± 1.89	94.15 ± 0.58	**96.69 ± 0.68**	93.96 ± 0.89	88.89 ± 0.0
	5	**92.09 ± 0.85**	91.98 ± 0.39	**92.09 ± 1.04**	85.2 ± 1.74	93.57 ± 0.58	**94.35 ± 0.34**	90.84 ± 0.89	87.52 ± 2.64
	10	**92.54 ± 0.9**	90.4 ± 1.28	90.17 ± 1.48	84.97 ± 1.93	**94.54 ± 0.89**	94.15 ± 1.55	91.81 ± 1.55	86.35 ± 3.33

Appendix D

Appendix: Modular prompt optimization for stance detection

Table A1: Sample distribution across train, validation, and test sets for each target in the stance detection datasets.

Dataset	Target	Train	Valid	Test	Total
COVID-19-Stance	Anthony Fauci	1,464	200	200	1,864
	Face Masks	1,307	200	200	1,707
	School Closures	790	200	200	1,190
	Stay at Home Orders	972	200	200	1,372
	Total	**4,533**	**800**	**800**	**6,133**
GunStance	guns should be banned	900	900	900	2,700
	guns should be regulated	936	932	932	2,800
	Total	**1,836**	**1,832**	**1,832**	**5,500**
SemEval-2016	Atheism	333	93	220	646
	Climate Change is a Real Concern	252	56	169	477
	Feminist Movement	427	97	285	809
	Hillary Clinton	442	115	295	852
	Legalization of Abortion	411	105	280	796
	Total	**1,865**	**466**	**1,249**	**3,580**
VaccineStance	Vaccines	407	578	1,015	2,000
	Total	**407**	**578**	**1,015**	**2,000**

D.1 Meta prompts

1. Instruction only update

Your task is to write only the {module_name} section of the prompt for stance detection towards the target {target}. Below are the previous prompts with their scores. The score ranges from 0 to 100, and higher scores indicate better quality prompts.

{prompt_1}

{score_1}

...

{prompt_n}

{score_n}

Carefully analyze the previous prompts and their scores, and write a new, improved prompt to replace the prompt below:

{to_rewrite_prompt}

Wrap the new, improved prompt between START and END tags. Do not explain your changes or include any other text.

2. Reasoning step

Your task is to point out the problems with the current prompt based on the wrong examples. The current prompt is:

{current_section_prompt}

But this prompt gets the following examples wrong. You should analyze the differences between wrong predictions and ground truth answers, and carefully consider why this prompt led to incorrect predictions.

Below are the examples with Tweets, Wrong predictions, and Ground truth answers.

{error_demonstrations}

Give a reason why the prompt could have gotten these examples wrong, and how we can improve the prompt? Wrap the reason with START and END tags.

3. Reasoning update

Your task is to write an improved prompt based on the feedback. Below is the current prompt with its score. The score ranges from 0 to 100, and a higher score indicates better quality.

Prompt: {current_prompt}

Score: {current_prompt_score}

Below are the problems with this prompt.

{problems_reasoning}

Carefully analyze the previous prompt and the problems with this prompt. Write a concise, new, improved prompt. Wrap the new improved prompt between START and END tags. Do NOT explain your changes or include any other text between START and END tags.

4. Per Example Reasoning

Your task is to point out the problems with the current prompt based on why this prompt leads to incorrect predictions. The current prompt is:

{`current_section_prompt`}

You should analyze the differences between wrong predictions and ground truth answers. Below is an example with a Tweet, a Wrong prediction, and the Ground truth answer.

Tweet: {`tweet`}

Target: {`target`}

Ground Truth: {`ground_truth`}

Wrong Prediction: {`wrong_prediction`}

Give a reason why the prompt could have gotten these examples wrong, and how we can improve the prompt? Wrap the reason with START and END tags.

5. Per example reasoning update

Your task is to write an improved prompt based on the feedback. Below is the current prompt:

{current_section_prompt}

Below is an example with a Tweet, a Wrong prediction, and the Ground truth answer.

Tweet: {tweet}

Target: {target}

Ground Truth: {ground_truth}

Wrong Prediction: {wrong_prediction}

Below are the problems with this prompt.

{problems_reasoning}

Carefully analyze the previous prompt and the problems with this prompt. Write a new concise and improved prompt. Wrap the new improved prompt between START and END tags. Do NOT explain your changes or include any other text between START and END tags.

D.2 Complete prompt

Example of a Complete Base Prompt for Stance Detection

1. Task Definition: Stance detection is a task of determining the attitude or position expressed by an author in a text regarding a particular topic or claim. Your task is to classify the stance of a tweet towards the target "`{target}`" as either Favor, Against, or Neutral.

2. Guidelines:

1. Stance Indicators: Look for keywords, phrases, emoticons, and tone that indicate a favor, against, or neutral attitude towards `{target}`.
2. Implicit vs. Explicit Stance: Consider both implicit and explicit expressions of stance. If a tweet implies a stance without directly stating it, use the context and language to determine the intended stance.
3. Sarcasm and Irony: Be aware of sarcasm and irony, which can be used to express an opposing stance. Look for inconsistencies between the literal meaning and the tone or language used.

3. Favor Definition:
Classify the tweet as favor if it:

1. Shows an explicitly positive stance: openly supports `{target}`
2. Uses pro-`{target}` keywords: languages that frames `{target}` in positive light

4. Against Definition:
Classify the tweet as against if it:

1. Shows an explicitly negative stance: openly opposes `{target}`
2. Uses against `{target}` keywords: languages that frames the target `{target}` in negative light

5. Neutral Definition:
Classify the tweet as neutral if it:

1. Shows no clear stance: it neither supports nor opposes `{target}`
2. Uses neutral or tentative language: e.g., simply asking a question or stating opinion that signals indifference/uncertainty

6. Output Format:
Strictly follow the guidelines above and classify the tweet if it is Favor, Against, or Neutral to the target "`{target}`". Reply with the classification label only, without any additional text or explanation.

Tweet: `{tweet}`
Stance: `[Favor/Against/Neutral]`

Figure A1: Example of a complete base prompt for stance detection. The complete prompt is broken down into 6 distinct modules.