# RETHINKING MISINFORMATION DETECTION: ACCOUNTING FOR CAREY'S VIEWS OF COMMUNICATION

**VANESSA BELANGER**

**Advisor:** Dr. André Vellino, Associate Professor
School of Information Studies

Thesis submitted to the University of Ottawa
in partial Fulfillment of the requirements for the
Master's Degree in Information Studies

School of Information Studies
Faculty of Arts
University of Ottawa

# Abstract

When evaluated on social media datasets, misinformation detection systems are typically assessed as if social media postings are a homogeneous communicative genre. This thesis challenges that assumption by arguing that the strong presence of mass-media content in widely used datasets conceals models' underperformance on real-world user-generated content. To address this issue, this study introduces a novel social media dataset with a labeling framework that distinguishes between news-generated and user-generated content. This allows for the first systematic comparison of language models' misinformation detection performance across communicative genres. Model performance is analyzed using two generalized linear mixed models to investigate main effects and interactions related to content type, domain, prompting strategy, and model architecture. The results reveal a consistent performance gap in which models generally perform better on news-generated content than on user-generated content. However, the magnitude of this difference varies across domains and training approaches.

# Acknowledgements

First and foremost, I would like to thank my supervisor, Dr. André Vellino, for his time, insight, and unwavering support throughout my research. I could not have completed this work without his continued enthusiasm and thoughtful feedback.

I am also deeply thankful to Professor Krisandra Ivings, whose Advanced Research Methods course played a pivotal role in refining and reframing this work. I am especially grateful for her generosity in supporting my academic progress.

I would also like to thank the May Court Club of Ottawa for awarding me the Susan Anderson Memorial Scholarship, which provided meaningful financial support during my second year of study.

I acknowledge the use of Claude (Anthropic) for assistance with debugging and refining scripts used in the experimental components of this research. Its use was limited to technical support; all ideas, methodological decisions, analyses, and interpretations are my own.

Lastly, a heartfelt thank you to my partner, Kat; our two cats, Suki and Ryo; and my family, who have been my pillars throughout this process. Your love, support, encouragement, and patience made this possible. I am also grateful to Emma, whose friendship was a much-needed source of levity during my experience in this program, reminding me to keep things in perspective.

# Table of Contents

# List of Figures

# List of Tables

# List of Abbreviations

**AAVE** African American Vernacular English

**AI** Artificial Intelligence

**AIC** Akaike Information Criterion

**AIO** Artificial Intelligence Overviews

**BERT** Bidirectional Encoder Representations from Transformers

**BERT-emo** BERT-model that integrates multiple sentiment (Emotion) features

**CLEAR** Concise, Logical, Explicit, Adaptive, and Reflective

**CNN** Convolutional Neural Network

**CoT** Chain of Thought

**COVID-19** Coronavirus Disease 2019

**DeBERTa** Decoding Enhanced BERT with Disentangled Attention

**DistilBERT** Distilled version of BERT

**DL** Deep Learning

**EANN** Event Adversarial Neural Network

**FakeBERT** Fake news detection in social media with a BERT-based approach

**FCC** Federal Communications Commission

**FND** Fake News Detection

**GLMM** Generalized Linear Mixed Model

**GPT** Generative Pre-trained Transformer

**IB** Information Behaviour

**LLAMA** Large Language Model Meta AI

**LLM** Large Language Model

**LM** Language Model

**ML** Machine Learning

**MoE** Mixture of Experts

**NEP** News Environment Perception

**NER** Named Entity Recognition

**NGC** News Generated Content

**NLP** Natural Language Processing

**NN** Neural Network

**RAEmoLLM** Retrieval Augmented LLMs Framework Based on Emotional Information

**RAG** Retrieval-Augmented Generation

**RNN** Recurrent Neural Network

**RLHF** Reinforcement Learning with Human Feedback

**RoBERTa** Robustly Optimized BERT Pretraining Approach

**SBERT-FC** Sentence Bert Fully Connected

**SHAP** Shapley Additive Explanations

**SLM** Small Language Model

**UGC** User Generated Content

**UNESCO** United Nations Educational, Scientific and Cultural Organization

**U.S.** United States

**WWW** World Wide Web

# 1. Introduction

Social media is a primary driver of information dissemination and a dominant way through which people encounter information online (Silva et al., 2021). While the accessibility of social media platforms has diminished the influence of traditional information institutions, it has also significantly increased the spread of misinformation (Shu et al., 2017). Misinformation refers to false information that is spread, regardless of the intention to mislead the receiver (Wu et al., 2019). The classification of false information depends on the agent's intention, the contents of the claim, and the context in which untruthful claims are disseminated. Due to this, 'Fake news' is often used as a unifying term to cover a wide variety of communication that is to some degree false (Pérez-Escolar et al., 2023).

The World Economic Forum (2024) recently identified misinformation dissemination as the biggest short-term global risk against social resilience. Social resilience refers to the capacity of individuals and communities to engage in and sustain positive social relationships while adapting to and recovering from stressors and experiencing social isolation (Cacioppo et al., 2011). As concerns about misinformation intensify, recent policy shifts by major platforms, most notably Meta and X, reflect a broader debate about the role of private companies in moderating online speech. Elon Musk, CEO of X, has argued that platform-level content filtering poses a threat to free expression, framing X as a neutral digital town square rather than an editorial authority (X, 2025). Similarly, Mark Zuckerberg, CEO of Meta, has articulated a renewed commitment to free speech, reducing the platform's reliance on third-party fact-checking in favour of a community-based flagging approach (Meta, 2025).

These developments come at an interesting time, as research indicates that malicious actors can use advances in artificial intelligence (AI) to pollute global information systems with false narratives (Zugecova et al., 2025). As misinformation becomes increasingly concise and appealing to receivers, its capacity to spread rapidly and influence public discourse has intensified (Guo et al., 2021). Therefore, addressing the erosion of public trust and restoring the reliability of digital information ecosystems has become an increasingly important challenge.

The consequences of misinformation extend beyond concerns about information quality and have tangible implications for democratic stability and public safety. In health contexts, a systematic literature review on health misinformation identified increased vaccine hesitancy, widespread misinterpretation of scientific evidence, negative mental health impacts, and the misallocation of health resources as being associated with exposure to misleading health-related claims (Borges do Nascimento et al., 2022). In political contexts, exposure to false or misleading news has been shown to influence citizens' beliefs during electoral processes, particularly when content aligns with partisan preference (Allcott & Gentzkow, 2017). Moreover, research has documented how states and political platforms use social media to shape public opinion at home and abroad, highlighting the political implications of information manipulation in digital environments (Bradshaw & Howard, 2018). In the context of war and violent conflict, research suggests that misleading information can facilitate conflict escalation, while its correction may contribute to peace-building efforts, signaling the role of information in shaping public narratives and exacerbating tensions (Lewandowsky et al., 2015). These contexts demonstrate that the circulation of misinformation online inhibits individuals' capacity for

informed judgment and weakens the foundation upon which social resilience and democratic governance depend.

## 1.1 Communicative functions of social media

Social media environments position individuals as active participants in the circulation of information online. On such platforms, participation serves as the dominant mode through which individuals recognize their community membership and inscribe themselves within social and political life (Barney et al., 2016). While some exchanges are purely meant to transmit messages to users, many social media interactions are participatory, emphasizing engagement and collective interpretation. These dynamics signify that communication on social media platforms serve purposes that extend beyond the transmission of information.

This thesis applies James Carey's (2009) ritual view of communication to better understand how different communication models are reflected in social media content (see Section 2.3). Carey introduced the distinction between transmission and ritual models of communication in an essay, which was later expanded into his book, *Communication as Culture: Essays on Media and Society* (2009). Working within American cultural studies and the broader field of media studies, Carey developed this framework as a critique of the dominant paradigm in communication research, which he argued had become overly focused on the mechanics of message transfer at the expense of understanding communication's cultural and community-building functions.

In the transmission model, communication is oriented toward accurate message transfer, emphasizing quality, timeliness, and clarity. In contrast, the ritual model frames communication as a participatory social process, through which shared meanings are reinforced, and community

boundaries are maintained (Carey, 2009). Within social media environments, content reflecting the values of these models of communication coexist. Posts by mass media news organizations prioritize the transmission of information, while posts by regular users function primarily as ritualized practices centred on participation and affiliation.

This distinction is particularly relevant for analyzing differences between news-generated content (NGC) and user-generated content (UGC). NGC is shaped by institutional norms of accuracy, timeliness of message transfer, and authority, closely aligning it with the values of the transmission model of communication. In contrast, UGC more often reflects ritual-oriented communication embedded within participatory social contexts. Recognizing these communicative functions as distinct instances of information on social media is important for examining how misinformation is created and circulated online and motivates the comparative analyses undertaken throughout this thesis (see Section 2.3.2 for further analysis of the framework's application to misinformation detection research).

## 1.2 Misinformation detection in digital contexts

Misinformation detection is a research area concerned with developing computational systems that effectively distinguish between accurate and misleading information in digital environments (Benny, 2024). Early approaches to misinformation detection relied primarily on supervised machine learning (ML) techniques that used linguistic features and metadata extracted from content. As research progressed, methods expanded to incorporate contextual information, such as temporal cues and user behaviour, as well as propagation-based signals that model how information spreads across social networks (Wu et al., 2019). More recent work has been shaped by advances in natural language processing (NLP), particularly the introduction of

deep learning (DL) and transformer-based language models (LMs). These models enable richer semantic representations of text and have substantially improved performance on misinformation detection benchmarks (Liu et al., 2024). As a result, transformer-based approaches are now the dominant paradigm in automated misinformation detection research.

Although these approaches differ in the types of signals they prioritize, they share a common assumption that misinformation can be identified through stable patterns in content, context, or diffusion. In practice, this has led to detection systems that treat misinformation as a largely uniform classification problem, removed from the communicative purposes and social settings in which content is produced and interpreted. Consequently, the underlying model of communication reflected within content—whether it is intended to inform, persuade, provoke, or reinforce group identity—often remains outside the scope of detection models. For datasets built on content scraped from social media, this entails treating material produced by news organizations, institutions, or public officials as equivalent to content generated by everyday users.

## 1.3 Problem statement, research rationale and approach

Despite advances in automated misinformation detection, current evaluation practices largely treat social media content as a homogeneous category, failing to account for differences in communicative function between UGC and NGC. As a result, detection systems are commonly assessed under assumptions that overlook how participatory and institutionally produced content differ in structure, purpose, and social context.

From the perspective of Carey's (2009) distinction between ritual and transmission models of communication, this represents a substantive limitation. Uniform detection strategies

may not be equally well-suited to content that serves distinct social and communicative roles. Given the absence of prior research using labeling schemes that distinguish between content types, it remains unclear whether automated detection systems perform consistently across content types that differ in origin and communicative purpose. This uncertainty limits our ability to holistically assess the real-world applicability of detection systems intended for social media platforms.

To address this limitation, the present study evaluates misinformation detection performance using a dataset that distinguishes between UGC and NGC sourced from social media platforms. Preserving this distinction enables direct comparison of model performance across content types aligned with different models of communication. In addition to content type, the evaluation spans multiple misinformation domains, including *Health*, *Politics*, and *War*. Models are assessed across all combinations of content type and domain, allowing for within-model comparisons of performance under varying communicative and topical conditions. This approach provides insight into whether and how differences in content source and communicative function may influence the effectiveness of automated detection systems, thereby addressing the gap identified in existing evaluation practices.

## 1.4 Research objectives

The primary objective of this research is to evaluate whether automated misinformation detection systems perform consistently across social media content that reflects ritual versus transmission models of communication. Specifically, the study aims to assess detection performance across content produced by news organizations and authoritative institutions, as well as content generated by everyday social media users.

A secondary objective is to examine how misinformation detection performance varies across domains characterized by distinct narrative conventions and social stakes. By jointly considering content type and domain, this research also seeks to determine whether communicative and topical contexts interact to influence the effectiveness of automated misinformation detection systems in real-world social media environments.

## 1.5 Research questions

**RQ1**: Do automated misinformation detection systems perform differently in classifying user-generated content (UGC) compared to news-generated content (NGC)?

**RQ2:** Does the performance gap between UGC and NGC vary across topical domains (health, politics, and war)?

**RQ3:** Are content type performance patterns consistent across prompting conditions (zero-shot vs. few-shot) and domain?

**RQ4:** Do content type and domain performance patterns differ between LLMs and fine-tuned BERT models?

## 1.6 Hypotheses

Based on Carey's (2009) ritual and transmission models of communication, as well as prior research on misinformation detection and LM performance, we formulated the following hypotheses:

- **H1:** Models will demonstrate higher performance on NGC compared to UGC.
- **H2**: Model performance will vary across domains (health, politics, and war).

- **H3:** Prompting strategy (zero-shot vs. few-shot) will influence model performance and moderate performance differences between content types
- **H4**: Performance across content type and domain will differ between LLMs and fine-tuned BERT-based models.

## 1.7 Thesis contributions

This thesis contributes to the study of automated misinformation detection in several ways. First, this thesis presents a novel application of Carey's (2009) ritual and transmission models of communication. By applying this framework to misinformation detection, we extend beyond a purely technical understanding of automated misinformation detection and achieve a more nuanced understanding of LM performance. Moreover, this thesis extends this contribution by evaluating the influence of content type on model performance, moderated by domain, prompting strategy, and model architecture. This approach enables a more nuanced assessment of model performance across multiple interacting factors.

Second, this research introduces a balanced dataset that distinguishes between UGC and NGC across multiple domains. To our knowledge, this represents a novel contribution that enables the first content-type-specific comparative analysis of misinformation detection system performance across content types with different communicative functions. The composition of the dataset itself also addresses several prevailing challenges in misinformation detection research, such as the scarcity of pre-labeled datasets, multi-domain datasets, and class-balanced datasets.

## 1.8 Thesis outline

This thesis begins by situating the state of misinformation on social media and introduces Carey's (2009) transmission and ritual models of communication as its guiding theoretical framework. Chapter Two reviews the literature on digital gatekeeping, recommender systems, information behaviour, machine learning, and automated misinformation detection. It draws on foundational and contemporary literature across these subjects to frame the need for an automated misinformation detection system that can be effectively deployed online, while also identifying current gaps in existing evaluation practices. Chapter Three outlines the study's methodological approach, including dataset construction, model selection, prompting strategies, and generalized linear mixed modeling (GLMM). Chapter Four presents our empirical findings and addresses differences in performance across content type, domains, prompting conditions, and model architectures. Chapter Five interprets these results through Carey's framework and discusses the implications for consistent and context-aware misinformation detection and dataset design. The thesis concludes by summarizing contributions, acknowledging limitations, and outlining directions for future research.

# 2. Literature review

Prior research on misinformation detection has largely approached the task as a technical filtering or gatekeeping problem, focusing on identifying and removing false content from digital platforms. This framing treats misinformation as a property of individual content items while failing to consider the models of communication reflected within different types of content, and the distinct roles they play within broader information ecosystems. Communication research suggests that communication is a ritual process through which people create shared meaning (Carey, 2009). From this perspective, different types of content should be understood as distinct forms of participation in meaning-making processes. Therefore, UGC and NGC may function differently within social networks, serving distinct purposes in terms of authority, social interaction, and meaning-making.

To address the research questions outlined in Section 1.5, this chapter reviews the literature on information gatekeeping, recommender system amplification, information-seeking behaviour, and automated misinformation detection. This review traces the evolution of information ecosystems, examining how recommender systems structure content exposure, how platform architectures mediate information circulation, and how users navigate these systems to find reliable information. The review then considers the emergence of LLMs, which have further complicated this landscape. Against this backdrop, the chapter identifies significant gaps in prior work, particularly the failure to account for different models of communication reflected within evaluated content and the limited attention to domain-specific evaluation methods. These gaps motivate this thesis's comparative analysis of LM performance across UGC and NGC domains,

investigating whether systems designed and evaluated without consideration for communicative context can reliably detect misinformation across varied forms of digital discourse.

## 2.1 Evolution of information gatekeeping

Gatekeeping is a central concept for understanding how information is filtered and disseminated within society. Traditionally, this process was mediated by human actors and institutional norms that governed the flow of information. However, the digital transformation of media has altered the nature of gatekeeping, shifting control from identifiable human actors and institutions to complex recommender systems such as social media feeds that prioritize content based on engagement metrics. This section traces the evolution of information gatekeeping from its origins in traditional media to its contemporary form in recommender systems. It also highlights how changes in gatekeeping mechanisms have reshaped the conditions under which we access and assess information, and how these changes have contributed to the proliferation of misinformation online.

### *2.1.1 Traditional gatekeeping theory and mechanisms*

In 1947, Kurt Lewin introduced the gatekeeping theory, which describes how information flows through decision-making points controlled by gatekeepers who determine what is disseminated and what is suppressed. Initially applied to food distribution, David Manning White (1950) applied the concept to mass communication and information science, where institutions such as libraries, news organizations, and academic publishers historically controlled the information the public could access (Shoemaker & Vos, 2009).

Libraries, for instance, played a crucial role in curating reliable information, ensuring that their collections upheld credibility, accuracy, and integrity (Moran & Morner, 2018). Similarly,

traditional news media employed a combination of editors and fact-checkers to maintain journalistic standards while determining which stories were considered newsworthy. Building on this recognition of editorial and curatorial authority, scholars began to study the decision-making processes that shaped the flow of information.

In one of the earliest gatekeeping studies, White (1950) conducted a case study of a "Mr. Gates" to explore how editors determined which stories from wire services to include in the daily paper. In his curatorial role, Mr. Gates rejected 90% of submitted stories and often let his personal preferences dictate which stories to include in the paper. As a result, the paper rarely published pieces on suicides and sensationalized news, and prioritized political news that aligned with Mr. Gates' own views (White, 1950). Although White's (1950) study demonstrated the importance of the individual processes that guide gatekeeping, a subsequent revaluation revealed that the types of stories that made it through to publication were directly proportional to the frequency with which the wire services provided them (Napoli, 2019). This reanalysis demonstrated that, in addition to Mr. Gates' own selections, a secondary level of gatekeeping by the wire services also played a key role in determining what got through.

### *2.1.2 Digital transformation of gatekeeping*

Using Hirsch's reanalysis, Napoli (2019) highlighted the significance of gatekeeper interactions in light of complex media systems. He stressed that, despite the internet initially being heralded for its neutrality and for undermining traditional intermediaries, contemporary media platforms introduce new, far more complex intermediaries (Napoli, 2019). For instance, the adoption of cable television led the Federal Communications Commission (FCC) to implement 'must carry' rules, which required local cable systems to provide local news

broadcasts to their consumers. Social media platforms, on the other hand, are not subject to any regulatory safeguards. As a result, the processes that determine what news is accessible to social media users are unclear (Napoli, 2019).

Underpinning this problem is algorithmic and ML gatekeeping, processes by which opaque automated systems filter, highlight, suppress, or otherwise play an editorial role in dictating information flow through online platforms (Tufekci, 2015). Early conceptualizations of algorithmic gatekeeping emphasized rule-based and procedurally defined systems (Diakopoulos, 2014). Today, however, recommender systems are driven by ML, enabling them to adapt dynamically based on user behaviour, engagement patterns, and large-scale data inference (Kugler, 2024).

Despite technological changes, the core gatekeeping operations described by Diakopoulos (2014) remain consistent across rule-based and ML-driven recommender systems. Most relevant to this thesis are the operations of prioritization and filtering. Prioritization emphasizes certain content through ranking mechanisms that embed value choices that determine what is pushed to the top of a feed (Diakopoulos, 2014). These criteria are often not publicly available, making it difficult to fully understand how content is ranked. Filtering involves including or excluding information according to defined rules, strongly shaping what appears in a user's feed (Diakopoulos, 2014).

On social media platforms, these two operations are at the core of recommendation systems (Kennedy, 2024). Although such systems can help users navigate an ever-expanding information environment (Kennedy, 2024), they have also played a central role in the

dissemination of misinformation and contribute to an increasing political polarization (Pariser, 2011; Vosoughi et al., 2018; Cinelli et al., 2021; Fernandez et al., 2024).

The absence of human editorial oversight on social media platforms enables the continuous circulation of unverified content. While this has expanded access to diverse forms of information, automated curation plays a decisive role in determining which content is most visible to individual users. Because these prioritize preference-aligned content and encourage passive information encounters, users may be especially vulnerable to misinformation encountered on social media. As social media's faciliatory role in the spread of false information becomes more influential, the development of automated misinformation detection systems for deployment on these platforms has emerged as a significant area of research.

## 2.2 Recommender system amplification of misinformation

Contemporary social media platforms rely on ML-driven recommender systems to curate content and deliver personalized user experiences (Kennedy, 2024). This section explores how these systems amplify misinformation—specifically, how cognitive bias, content-prioritization logic, and the formation of echo chambers and filter bubbles contribute to its spread.

### *2.2.1 Bias in content prioritization*

Platforms that use ML-driven recommender systems at scale have been widely criticized for failing to adequately account for the adverse consequences of their content curation criteria (Pariser, 2011). As platforms have grown more influential, their social and political repercussions have intensified, particularly with a shift among leading platforms toward promoting popular content on user feeds regardless of whether users follow their source accounts. Twitter (now X) was the first major platform to implement this approach. This move

drew substantial scrutiny for undermining users' autonomy over their feeds and for amplifying political rhetoric and misinformation (Darcy, 2019).

In response to these concerns, Fernandez et al. (2024) sought to examine the role of recommendation systems in the spread of misinformation on Twitter. They found that these systems suffer from popularity bias and that systems designed to mimic popularity metrics were more likely to boost false information. In light of these findings, their study explored pre- and post-model interventions, testing collaborative filtering and re-ranking strategies to reduce bias and limit the spread of misinformation (Fernandez et al., 2024). Their results suggest that collaborative filtering techniques are a viable approach for limiting the spread of misinformation by recommender systems. Despite these suggestions, leading social media platforms continue to prioritize engagement over informational integrity and user satisfaction (Milli et al., 2025; Avram et al., 2020). A preregistered audit of a X's ranking system by Milli et al. (2025) found that it amplifies emotionally charged and hostile content, even when users report dissatisfaction with such outputs. They found that while incorporating user-stated preferences reduced exposure to angry, partisan, and out-group-hostile content, the approach risks reinforcing filter bubbles, highlighting the persistent tension between personalization and content diversity.

### *2.2.2 Echo chambers and filter bubbles*

Research suggests that ML-driven content curation contributes to ideological isolation, reinforcing user biases rather than exposing them to diverse perspectives (Cinelli et al., 2021). Pariser (2011) coined the term "filter bubble" to describe how personalized recommender systems create isolated information environments that limit users' exposure to opposing viewpoints. Similarly, Cinelli et al. (2021) examined the "echo chamber effect" across major

social media platforms. Their comparative analysis of over 100 million pieces of content found that polarizing topics are overrepresented in homophilic clusters—online communities where users predominantly interact with like-minded individuals. Their findings revealed that Facebook and Twitter are particularly prone to echo chambers, with highly polarized topics correlating with increased misinformation spread (Cinelli et al., 2021). These studies highlight the role of automated gatekeeping mechanisms in shaping public perception and the need for interventions to counteract biased content amplification.

### *2.2.3 Human factors in misinformation spread*

Beyond system-level bias, human cognitive bias also contributes to the spread of misinformation. Vosoughi et al. (2018) conducted a large-scale study of 126,000 Twitter news stories, finding that false news spreads farther and faster than accurate news, particularly in political contexts. They found that false tweets were 70% more likely to be retweeted than true ones, and that true tweets took 20 times as long to reach a cascading depth of 10.

They also found that humans, rather than bots, drive the majority of misinformation spread, challenging assumptions that automated systems are the primary culprits. People are more likely to share novel, emotionally provocative content over fact-based information, making engagement-optimized ranking systems that mimic popularity even more susceptible to amplifying misleading content (Fernandez et al., 2024). Findings by Avram et al. (2020) also point to the role of cognitive bias in the spread of misinformation. Using a news literacy game that simulates a social media feed, they found a positive correlation between higher engagement and the likelihood that users would share questionable content rather than fact-check it. Their

findings suggest that social metrics strongly influence user interaction with low-credibility information (Avram et al., 2020).

However, these human factors do not operate uniformly across all content types. In a study examining the persuasive effectiveness of health messages across different communication channels, Lee et al. (2022) found that users engage differently with professionally produced news content than with peer-generated social media posts. Specifically, they found that participants suspected fewer ulterior motives when reading health-related content presented as a Facebook post than when it was presented as a news article, suggesting that users apply different heuristics when evaluating UGC versus NGC (Lee et al., 2022).

Similarly, a study by Zhou (2025) comparing perceptions of credibility, bias, and the impact of citizen versus traditional journalism found distinct perceptual differences across the two sources. For instance, participants in Zhou's (2025) study perceived citizen journalism with verified sources as more credible than traditional journalism with unverified content. However, in contrast to Lee et al.'s (2022) study, participants in Zhou's (2025) study perceived citizen journalism as more biased than traditional journalism covering the same events. These findings indicate that users apply different evaluative frameworks to content source types across contexts, which carries implications for how misinformation is processed and shared online.

Moreover, a study by Boot et al. (2021) revealed that peer-user comments and social cues largely influence the processing and evaluation of NGC on social media. They found that negative comments affected users' personal opinions of content, thereby lowering intent to share, reducing agreement with ideas, and decreasing perceptions of content credibility (Boot et al., 2021). These findings suggest that NGC on social media operates within ritualistic meaning-

making processes, in which a sense of belonging and peer interaction shape evaluative processes distinct from those in traditional news consumption.

From Carey's (2009) ritual view of communication (see Section 2.3), these evaluative processes reflect the significance of community in constructing shared meaning. If users encode and respond to different cues when engaging with UGC versus NGC, then these content types likely exhibit distinct linguistic and structural features that LMs may learn to recognize. This could lead to performance differences in misinformation detection across UGC and NGC. The interaction between cognitive biases, recommender system amplification, and content-type-specific information behaviours creates a context-dependent vulnerability to misinformation, motivating this study's exploration of whether detection performance differs between UGC and NGC (RQ1, RQ2) and across misinformation domains (RQ4).

## 2.3 Conceptual framework

This section describes James Carey's (2009) ritual view of communication framework, which serves as the conceptual anchor for interpreting this study's empirical findings. Carey's framework provides a lens for understanding how content generated by news organizations (NGC) and everyday social media users (UGC) differ in communicative purpose and social function, and why that may manifest as measurable differences in model performance. The section first describes Carey's conceptualization of transmission and ritual models of communication, outlining their origins and core characteristics. It then explains how this distinction maps onto NGC and UGC in social media environments, where news organizations operate as institutional information brokers while users engage in community-oriented, identity-affirming communication practices. Finally, it demonstrates how this framework is useful for

misinformation detection research, enabling a more nuanced understanding of detection system performance and bias.

### *2.3.1 Carey's views of communication*

Carey (2009) theorized that the transmission view of communication stems from early desires to optimize the speed and quality of message transfer as it travelled through space and time via early communication technologies such as the telegraph and telephone. He suggested that this focus is why communication and information became entangled with transmission itself. The transmission view, which is rooted in Claude Shannon's (1948) theory of communication, conceptualizes communication as the extension of messages across geography for the purposes of control and influence. This view dominated early communication studies and continues to shape how we think of information flow, particularly in digital contexts.

To contrast this view, Carey (2009) presented the ritual view of communication. In this view, communication serves to maintain society in time and create an ordered, meaningful cultural world. As such, to partake in communication is to be part of a community. In an earlier publication, he illustrates this point by describing engagement with newspapers, stating that the practice is less about "sending or gaining information" and is rather about "attending a mass, a situation in which nothing new is learned but in which a particular view of the world is portrayed and confirmed…" (Carey & Adam, 2008, p. 16). Carey's original intention was to recover an understanding of communication as a symbolic process whereby reality is produced, maintained, repaired, and transformed. He drew on cultural anthropology and the work of scholars like John Dewey to argue that communication is fundamentally about the construction and maintenance of shared meaning within communities

### *2.3.2 Application to misinformation detection research*

Distinguishing between several types of false information requires the epistemological assumption that information carries intention—where dimensions such as the agent's intention, the claim's content, and the context in which untruthful claims are disseminated ultimately determine their classification (Wardle & Derakshan, 2017). For instance, misinformation is distinguished from disinformation by the absence of an agent's intention to spread false information to mislead the receiver (Wu et al., 2019). Bastos and Tuters (2023) suggest that the epistemological assumption that false information possesses fixed, distinct characteristics partially explains why many misinformation models ground themselves in ideas of transmission and control. Examples include contagion modeling in media effects research (Lerman & Ghosh, 2010) and fact-checking initiatives designed to flag or remove false information. These approaches value interrupting transmission pathways over understanding the cultural work that mis/disinformation performs within communities.

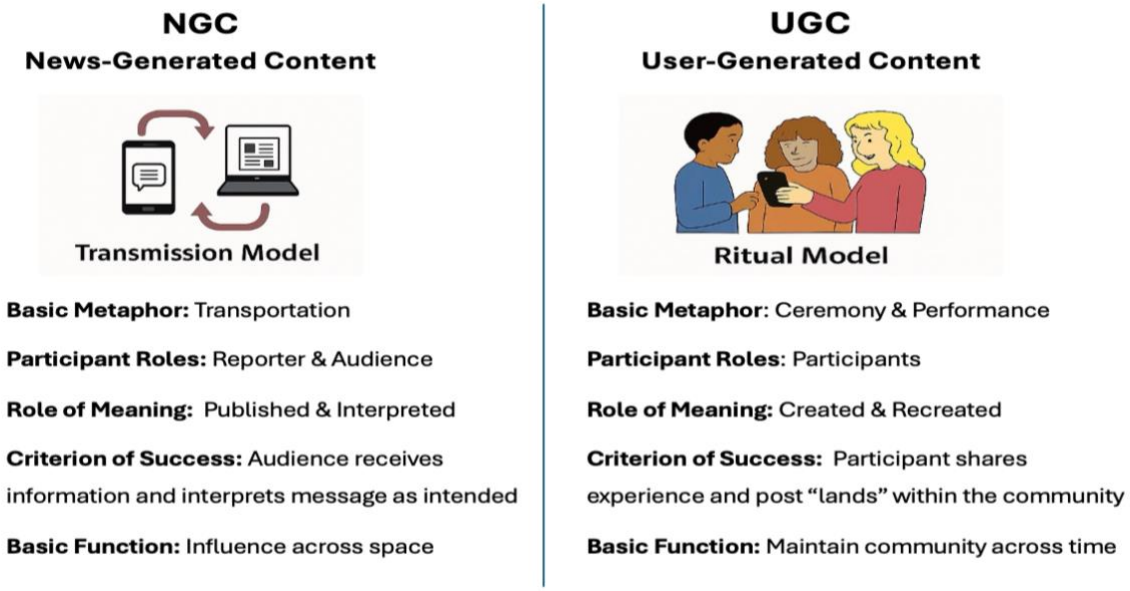
This study's distinction between NGC and UGC mirrors Carey's transmission-ritual dichotomy. NGC conforms to standardized journalistic conventions that emphasize clarity, structure, and epistemic authority, reflecting the transmission model of communication's emphasis on accurate message transfer and institutional control. However, this institutional authority is itself ritualistic, as the credibility of news organizations depends on shared beliefs about journalistic objectivity and professional standards, maintained through ritual practices of citation and appeals to expertise.

In contrast, UGC serves a ritual function to maintain community boundaries, perform social identity, and reinforce shared worldviews. Platform-specific vernacular, social affiliations,

and various informal communication styles shape UGC's presentation. For instance, in a focus group study of frequent multi-platform social media users, the Pew Research Center (2023) found that posting decisions were determined by the platform itself and the anticipated audience. Similarly, Alhabash and Ma (2017) found that motivations for social media use differed across platforms, with participants reporting that information sharing drove Facebook use, while Instagram and Snapchat served purposes for self-expression and documentation (Alhabash & Ma, 2017).

**Figure 1**

*Views of Communication in Online Content*




*Note.* Adapted from James Carey's (2009) ritual view of communication dichotomy, applied to forms of online media content.

These platform-specific practices exemplify Carey's (2009) claim that to partake in communication is to be part of a community. In ritual communication, effectiveness is measured

by social cohesion rather than informational accuracy or clarity. This ritual view produces linguistic features that deviate from journalistic conventions. Platform-specific vernacular, implicit cultural references, and context-dependent meaning are elements that enable community participation and boundary maintenance. For detection systems trained on transmission-oriented NGC, these features may be misinterpreted as signals of unreliability.

Sap et al.'s (2019) research on automated hate-speech detection demonstrates this problem. Their findings show that linguistic markers such as African American Vernacular English (AAVE) are flagged at disproportionately higher rates by automated systems (Sap et al., 2019). If LMs associate certain registers with reliability, they may misclassify ritual communication—particularly communicative styles used by marginalized communities—as false.

Current misinformation detection systems face a fundamental problem: detection systems are trained and evaluated on datasets that treat content as possessing a single, unified communicative orientation. This research addresses this gap by using Carey's (2009) framework to compare model performance across NGC and UGC. The quantitative analysis performed in this study reveals whether the transmission-ritual distinction produces measurable differences in detection capabilities. This study uses a labeling schema that differentiates content according to the communicative model they reflect, in addition to domain and veracity.

By demonstrating that misinformation detection requires theoretical frameworks that account for power, community, and communicative purpose, this thesis challenges the notion that improved algorithms and datasets serve as optimal, neutral technical solutions. Rather, it demonstrates that the technical challenge of misinformation detection cannot be separated from

questions about what information does in social contexts and whose ways of communicating are valued or rendered suspect.

## 2.4 Information-seeking in the digital age

To fully understand the conditions that contribute to misinformation susceptibility, we must also examine how users encounter, process, and share information within ML-driven curated environments. Wilson's (1981) Information Behaviour (IB) theory provides a foundational lens for understanding user-level processes, particularly as they relate to this thesis's research questions regarding misinformation detection on social media. While such frameworks have been central to shaping our understanding of how individuals seek and process information, the following section argues that some may inadequately account for contemporary social media environments. On social media, traditional evaluative behaviours are often bypassed as users consume information incidentally rather than through intentional search. Further complicating approaches to misinformation detection, the literature in this section reveals that information behaviours vary substantially across different social media domains, where the social and communicative functions of UGC can amplify misinformation risks.

### *2.4.1 Evolution of information-seeking models*

Information science has long examined how individuals search for, encounter, and process information. As technology advances, so do the mechanisms for seeking information. Traditional models of information-seeking behaviour provide important insights, but the rise of social media and AI-driven search systems has introduced new challenges and opportunities for research.

Classic models of information-seeking behaviour conceptualized search as a structured, goal-oriented process. Wilson's (1981) IB model suggests that information-seeking arises once the user identifies a need for information. Once identified, the user performs several actions and works through stages to satisfy the information need. Ellis (1989) described several core sub-processes, including starting (initiating a search), chaining (following references), browsing (exploring related topics), monitoring (tracking new information), verifying (assessing accuracy), and ending (attempts to conclude the process through a final search). Norman's (1988) executive evaluation model distinguishes execution (search actions) from evaluation (assessing retrieved information), pointing to the significance of action in goal-oriented search. Bates' (1989) berry-picking model challenged the notion of linear information retrieval by arguing that information needs evolve dynamically, with users modifying their queries as they discover new information.

Although these models constitute the foundational theoretical frameworks for understanding information-seeking behaviour, our review of the literature in Section 2.2 suggests that their applicability to contemporary media environments is limited. The search models proposed by Bates (1989), Ellis (1989), Norman (1988), and Wilson (1981) presuppose that individuals engage with information seeking as an intentional, goal-oriented, and evaluative process. However, social media and algorithm-driven content delivery facilitate passive encounters with information rather than active search efforts.

Kulthau's (1991) Information Search Process (ISP) model offers a holistic representation of the search process by recognizing that both the cognitive and affective dimensions experienced by the searcher shape the process of filling knowledge gaps. Kulthau's (1991) recognition of affective responses is particularly relevant to modern research on emotionally

charged content and its role in amplifying engagement and dissemination on social media platforms (Pariser, 2011; Vosoughi et al., 2018; Cinelli et al., 2020; Fernandez et al., 2024). However, other dimensions of the ISP model are less compatible with the realities of system-curated environments. For instance, Kulthau (1991) identifies an "exploration" stage during which individuals typically encounter information that challenges or conflicts with their prior assumptions. Yet, accounts of filter bubbles and echo chambers (Pariser, 2011; Cinelli et al., 2020) suggest that personalized feeds reduce the likelihood of such encounters. Furthermore, the ISP model assumes a sequence of intentional and reflective activities such as documentation and assessment, which contrast with the largely passive and incidental nature of information exposure on social media.

Rather than focusing on the strategies and stages involved in the search process, Brenda Dervin's (1998) sense-making focuses on the cognitive and social processes through which people navigate uncertainty and construct understanding. Dervin conceptualizes sense-making as a process of bridging knowledge gaps and integrating new information into pre-existing cognitive schemata (Pentina & Tarafdar, 2014).

Mirbabaie and Zaptaka's (2017) paper on sense-making in social media crisis communication demonstrates the relevance of sense-making theory to information-seeking on social media. They summarize several research papers on crisis communication that, together, demonstrate how users engage in information-sharing to collectively make sense of developing situations (Mirbabaie & Zapatka, 2017). For instance, during the 2011 Egyptian revolution, hashtags served as mechanisms for gathering information, raising situational awareness, and providing updates, thereby supporting collective sense-making (Oh et al., 2015). Similarly, an

analysis of U.S. campus shootings revealed that users engage in intensive information sharing by using topic-related hashtags to facilitate both individual and collective sense-making (Heverin & Zach, 2012).

Sense-making on social media involves integrating information from multiple sources, including both official, trusted sources and UGC (Schafer et al., 2007; Oh et al., 2013). The communication rituals through which sense-making occurs, such as asking questions, sharing interpretations, and seeking community validation, produce discourse patterns distinct from transmission-oriented information delivery characteristic of NGC. As knowledge gaps are bridged, users shift from prioritizing information gathering to developing and communicating their own opinions. At this stage, opinion-related tweets tend to increase (Heverin & Zach, 2012; Vos & Buckner, 2015), and the difference between UGC and NGC becomes pronounced. These distinctions have important implications for automated misinformation detection on social media platforms.

### *2.4.2 Domain-specific misinformation challenges*

Using social media as a primary source of information poses many risks, but some areas carry greater consequences than others. As seen during the 2016 U.S. presidential election's subsequent post-truth era (Lewandowsky, 2017) and the COVID-19 pandemic, misinformation has significant real-world implications for health and political discourse.

Thackery et al. (2013) surveyed 1745 young adults to establish the frequency of various types of online health-seeking behaviours. In their survey, 41% of respondents reported using social media to seek health information, and the likelihood of turning to social media for health information doubled among those diagnosed with a chronic disease (Thackery et al., 2013).

Similarly, Kishimoto and Fukushima (2011) found that out of a sample of 4861 Japanese consumers, 51% reported obtaining information related to drugs on social media, most notably anonymous web communities. Another study investigating where college students seek health-related information online found that 32.8% of consumers between 18 and 30 reported using social media as an information source (Prybutok & Ryan, 2015). A commonly reported reason is the prevalence of others having a similar issue, highlighting the importance of the social dimension of information seeking through social media.

After the COVID-19 pandemic, Neely et al. (2021) conducted a sentiment analysis of online health information. They found that despite widespread mistrust in online health information, users continue to rely on social media for pandemic updates. In their study, 76% of respondents reported relying on social media at least "a little", and despite mistrust, 63.6% of users reported they were unlikely to fact-check information encountered online with a credible information source (Neely et al., 2021), illustrating the paradox of high engagement with low-trust content.

Although we are seeing rising numbers of people using social media for news consumption (Pew Research Center, 2024), research indicates a decline in trust in political content on social media. For example, Morris and Morris (2023) investigated whether Americans were reducing their use of social media to consume political content and found an increase in distrust in information found on social media between 2016 and 2019. These findings were consistent with a 2023 UNESCO report, which found that 68% of respondents identified social media as the primary driver of disinformation (UNESCO, 2023). Although research indicates

increased skepticism toward online content (Neely et al., 2021; Morris & Morris, 2023), Park et al. (2023) found that those who continue to rely on social media as a primary information source are less likely to seek information elsewhere.

Together, these findings highlight domain-specific challenges in misinformation research. Although users may exhibit heightened skepticism towards health content encountered online (Neely et al., 2021), research points to the role of the social dynamics such as connection, support, and community that are integral to social media and often provide comfort for users seeking health related information (Kishimoto & Fukushima, 2011; Thackery et al., 2013; Prybuto & Ryan, 2015). Informational vacuums reinforced by filter bubbles and echo chambers (Pariser, 2011; Cinelli et al., 2020) become all the more effective in this context, as reliance on social media discourages recourse to alternative sources (Park et al., 2023), thereby narrowing users' exploration for credible resources and entrenching the conditions under which misinformation can thrive.

## 2.5 LLMS in information retrieval and misinformation

The integration of LLMs into information retrieval systems has changed how users access and evaluate online content (Chapekis & Lieb, 2025). While LLMs offer powerful capabilities for analyzing and classifying text at scale, their deployment in misinformation detection raises important questions about reliability, bias, and performance consistency across diverse content types. This section examines the current state of LLM integration into information ecosystems, beginning with their role in reshaping information retrieval practices and user behaviour. It then turns to research on LLM limitations, biases, and task-dependent performance, highlighting factors that may influence their effectiveness when applied to misinformation detection.

Understanding these capabilities and limitations is essential for evaluating whether LLM-based detection systems perform reliably across content that reflect different models of communication.

### *2.5.1 LLMs and the transformation of information retrieval*

The evolution of search engines in the 1990s shaped interactions with the World Wide Web (WWW). The earliest system, *Archie* (Emtage, 1990), indexed file directories across online servers, enabling users to locate the systems hosting desired files. This was followed by platforms such as *Yahoo! Search,* whose hierarchical directory structure established it as the first widely adopted search engine, and *AltaVista*, which introduced crawler-based indexing and supported natural language queries before being absorbed into Yahoo. *Google* (Page et al., 1998) set a new standard for relevance ranking with their PageRank algorithm, rapidly establishing itself as the dominant search engine and rendering its name synonymous with online searching. Since then, natural language querying has remained the dominant paradigm for information retrieval on the web.

In 2024, the integration of generative AI into search engines marked a new era in online information retrieval. Elizabeth Reid, Vice President and Head of Search at Google, described the company's Gemini model as taking the "legwork" out of searching by generating *AI Overviews* (AIO)*,* summaries that compile relevant information at the top of results pages (Google, 2024). According to Reid (2024), these summaries encouraged users to visit a broader range of websites by embedding source links. However, recent findings directly challenge this claim. A Pew Research Center study of almost 69k Google searches found that users exposed to AIOs were significantly less likely to click external links, with only 8% of searches resulting in a click compared to 15% without summaries (Pew Research Center, 2025). They also found that

sessions were more likely to terminate on pages with AI summaries (26%) than without (16%). However, the findings also indicate that Google AIOs mirror standard search results, with Wikipedia, YouTube, and Reddit accounting for 15% of the results listed in summaries. As of March 2025, approximately one-in-five Google searches produce an AIO (Pew Research Center, 2025). An analysis by Authoritas (2025) found that a site previously ranked first in a search result could lose approximately 79% of its traffic for that query if results are delivered below an AIO. Beyond traffic implications, this data illustrates a broader transformation in how users interact with information online.

The shift toward LLM-mediated information retrieval has profound implications for the proliferation of misinformation online. While traditional search required users to evaluate multiple sources, AIOs present synthesized conclusions that users increasingly accept without verification (Pew Research Center, 2025). This change diverges from classical information-seeking models, where encountering multiple sources creates opportunities for cross-validation and critical assessment. Instead, when AIOs present synthesized conclusions as authoritative responses, users engage in what Carey (2009) might characterize as a ritual affirmation of the technology's epistemic authority. The social act of “googling” thus shifts from information-seeking behaviour to a form of technological communion, where the algorithm's output is trusted simply because consulting it has become a culturally accepted form of knowledge verification.

This ritualistic dimension is particularly consequential for misinformation detection. As users increasingly defer to automated synthesis without engaging source material, the question of whether detection systems can reliably distinguish between accurate and misleading content becomes critical. If AIOs and other LLM-mediated systems are to serve as gatekeepers of

information accuracy, understanding their performance across the diverse content encountered in social media environments is essential.

### *2.5.2 LLMs as information retrieval tools*

Before evaluating LLMs as misinformation detection tools, it is important to understand their documented limitations, biases, and the ways users interact with their outputs. A growing body of research examines how LLMs perform in information-processing contexts, revealing patterns of overconfidence, task-dependent accuracy and embedded biases that have implications for their deployment in automated detection systems (Pennycook et al., 2020; Choi et al., 2024; Spatharioti et al., 2025).

Spatharioti et al. (2025) found that LLM-based information retrieval required roughly half the time of traditional search engines, with users reaching decisions more quickly and with fewer steps. However, accuracy was dependent on task complexity. While performance in both conditions was similar for simple tasks, significant drops in accuracy occurred only in LLM-mediated contexts as tasks became more difficult. Overreliance was evident, as 60% of participants in the LLM condition issued only a single query before making a final decision on a task categorized as difficult (Spatharioti et al., 2025). Choi et al. (2024) observed a similar effect, finding that LLMs can induce an anchoring bias, leading users to trust AI-generated responses and to disproportionately discount their own judgment. These findings suggest that LLM performance may degrade on complex or nuanced content. This consideration is directly relevant when evaluating their effectiveness in detecting misinformation across domains with varying levels of narrative complexity.

Research also indicates that reliance on LLM outputs may persist even when alternatives exist. Bogert et al. (2021) found a positive correlation between increased task difficulty and participants' likelihood of aligning with AI recommendations rather than peer advice. This contrasts with an earlier study by Logg et al. (2019), who reported that professionals were less likely than non-professionals to trust algorithmic advice, even when it improved their performance. The discrepancy between these studies suggests that such skepticism may gradually erode as LLMs become more sophisticated and widely adopted, a trend that demonstrates the importance of rigorously evaluating detection system performance before deployment.

The persuasive capacity of LLMs further complicates their role in misinformation contexts. Costello et al. (2024) examined whether conversational AI could reduce conspiratorial beliefs through personalized dialogue. Their study found that after three rounds of conversation with an LLM designed to reduce participants' belief in their chosen conspiracy theory, average belief decreased by 20% (Costello et al., 2024). A professional fact-checker subsequently verified that 99.2% of the claims used by the LLM were true, suggesting that the model's persuasive strategies were grounded in factual information. Notably, the dialogues also reduced belief in unrelated conspiracy theories, demonstrating the broader impact of LLM-based persuasion. While these findings illustrate the potential of LLMs to mitigate misinformation, they also reveal the ethical risks associated with their persuasive capabilities. If deployed without sufficient oversight, the same authoritative tone that enables belief correction could obscure classification errors in detection systems, leading users to accept false negatives or false positives as accurate assessments.

### *2.5.3 Risks and limitations of LLM use in detection systems*

The risks associated with LLM deployment in misinformation detection extend beyond user behaviour, beginning with biases embedded in the models themselves. Fang et al. (2024) evaluated seven leading LLMs by prompting them to generate news articles from headlines, finding substantial gender and racial bias, with notable discrimination against females and Black individuals. These findings align with broader concerns that LLMs may reflect and amplify societal biases unless explicitly mitigated.

Prompt sensitivity further complicates the deployment of LLMs in detection tasks. Fernández-Pichel et al. (2024) compared LLMs and search engines for health-related queries, finding that while LLMs often outperformed search engines, their responses were highly sensitive to phrasing. Even with carefully crafted prompts, LLMs produced authoritative-sounding but incorrect outputs, exposing persistent knowledge gaps. This finding is directly relevant to misinformation detection, where variations in how claims are framed, the linguistic style of UGC versus NGC, or domain-specific terminology may influence model predictions. Coupled with findings from the Pew Research Center (2025), which indicate that users are less likely to consult additional sources when presented with LLM outputs, the risk of undetected classification errors becomes pronounced.

Promising directions have nonetheless emerged. Fernández-Pichel et al. (2024) found that retrieval-augmentation, a technique that enables models to draw on live web content, significantly improves the accuracy of smaller models by bridging knowledge gaps. Mannuru et al. (2024) advocated for a hybrid approach, in which LLMs complement rather than replace search engines. They emphasize LLMs' use for nuanced queries, summarizing long documents, and answering

simple factual questions. However, they caution that given limitations in accuracy, outdated knowledge, and weak sourcing, human-directed evaluation with AI-driven augmentation remains the most reliable way to reduce overreliance and vulnerability (Mannuru et al., 2024).

These findings reveal the nuances of integrating LLMs into the digital misinformation landscape. While they offer powerful capabilities for automated detection, they risk reproducing and obscuring misinformation if deployed without sufficient guardrails. The efficacy of LLMs in misinformation detection hinges on model performance and transparency about their limitations, awareness of embedded biases, and recognition that performance may vary across content types and domains. This study builds on these insights by examining how LLM-based detection systems perform across UGC and NGC, identifying whether models exhibit differential accuracy across misinformation domains, and determining which content characteristics most frequently elicit misclassifications. By preserving the distinction between content types that reflect different models of communication, this research addresses whether current detection approaches adequately account for the diversity of content encountered in real-world social media environments.

## 2.6 Technical foundations for misinformation detection

The emergence of LLMs represents a critical turning point in the evolution of NLP. This development enabled machines to perform increasingly sophisticated tasks, such as question answering, summarization, and problem-solving, at unprecedented speeds. While the scale and scope of modern LLMs have opened new frontiers in automated content evaluation, they also demand renewed scrutiny of the linguistic, cognitive, and social assumptions embedded in their architectures and applications. This section situates LLMs within the historical arc of NLP

development, outlines the key language-processing techniques relevant to misinformation detection, and critically assesses their limitations when applied to real-world, domain-diverse information environments. Of particular concern is whether detection systems trained predominantly on structured, institutionally produced content perform reliably when applied to the informal, context-dependent discourse characteristic of UGC on social media platforms.

### *2.6.1 Evolution of NLP and LLMs*

Early NLP models relied on rule-based systems, in which linguistic rules and patterns were manually encoded to parse and generate text. These systems were inflexible and ill-equipped to handle the variability of natural language. Statistical models in the mid-20th century, including n-gram approaches (Shannon, 1948) and Hidden Markov Models (Baum & Petre, 1966), introduced probabilistic reasoning, allowing systems to learn from data patterns. However, these approaches were limited by their inability to model long-range dependencies and contextual meaning.

McCulloch and Pitts' (1944) neural network is credited with the idea of modelling a computer based on the human brain, but the technological limitations of the time prevented its fruition. Rumelhart et al.'s (1986) work on backpropagation, a training algorithm for neural networks that adjusts model weights to minimize error, laid the foundation for the deep learning revolution. Subsequent advances, such as Recurrent Neural Networks (RNNs) designed to capture dependencies in sequential data, and Convolutional Neural Networks (CNNs), specialized for detecting local patterns through convolutional filters, expanded the horizons of NLP by enabling more effective handling of sequence structures and spatial relationships in text (Eisenstein, 2019, p.13).

The introduction of the Transformer architecture (Vaswani et al., 2017), built on self-attention mechanisms, marked the next paradigm shift. Transformers overcame the bottlenecks of earlier architectures and enabled the pre-training of large-scale LLMs on massive corpora, enabling them to achieve broad generalizability across tasks. Crucially, transformer-based models introduced transfer learning to NLP, enabling pre-trained models to be fine-tuned for a wide variety of downstream tasks. This shift not only boosted performance but also made powerful NLP tools broadly accessible. Models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLNet (Yang, 2019), T5 (Raffel et al., 2020), and GPT-3 (Brown et al., 2020) exemplify this breakthrough, each representing a significant milestone in the evolution of LLM research.

Transfer learning has profound implications for misinformation detection. Pre-trained models can be fine-tuned on labeled misinformation datasets, reducing the need for massive domain-specific corpora. However, this approach also introduces a critical complication that is of utmost relevance to this thesis. Through transfer-learning, model performance is shaped not only by the fine-tuning data but also by the characteristics of the pre-training corpus. If pre-training corpora are dominated by formal, edited text, models may develop representations optimized for transmission-oriented content while underperforming on ritual-oriented, participatory content such as UGC. This representational gap motivates the present study's focus on evaluating whether detection systems exhibit differential performance across content types that reflect these distinct communicative functions.

### *2.6.2 Core NLP tasks for misinformation detection*

Misinformation detection tasks in NLP are typically framed as classification pipelines, structured, end-to-end automated workflows that transform raw input data into specific, predefined class predictions (IBM, n.d.). In text classification, models receive text as input and are tasked with correctly assigning labels to each document or instance (Eisenstein, 2019, p. 13). However, the effectiveness of these classification approaches, depends heavily on the linguistic features available in the input text and the extent to which those features generalize across different content types.

Sentiment analysis, a subset of text classification, involves determining the sentiment or opinion of polarity of a given text input (Eisenstein, 2019, p. 67). In direct bag-of-words text classification, a classifier training method that focuses on word frequency rather than word order and grammar, sentiment analysis can involve models that rely on sentiment lexicons and label text as positive, negative, or neutral in polarity (Eisenstein, 2019, p. 67). This approach is widely considered ineffective for short text documents that contain hypothetical or nonfactual descriptions of events (Eisenstein, 2019, p. 68). Despite this limitation, sentiment analysis remains both a core detection strategy and a supplementary feature to larger detection frameworks. Early studies found that fake news often exhibits stronger negative sentiment than credible information (Dey et al., 2018; Yang et al., 2019) and that sentiment-aware classifiers can outperform their baselines using only text features (Ajao et al., 2019).

Recent approaches highlight the role of sentiment in both news content and user interactions. For example, Zhang et al. (2021) introduce the concept of "dual emotion," an approach that models both the publisher's sentiment in the news-text and the emotions elicited

by users in the comments. Their findings reveal that a strong divergence between these two groups' emotional signals can be indicative of misinformation (Zhang et al., 2021)

These findings are especially relevant to the present study's distinction between UGC and NGC. UGC, embedded within participatory social contexts, may exhibit more diverse and polarized sentiment patterns than institutionally produced news content. If sentiment-based features are weighted heavily in detection models, performance discrepancies between UGC and NGC may emerge, particularly in emotionally charged domains such as Health or Politics. Understanding whether sentiment plays a differential role across content types informs the interpretation of model performance patterns observed in this thesis.

In information extraction, models construct a knowledge base by processing natural language text and using embedded knowledge to complete a specific task (Eisenstein, 2019, p. 379). Within this paradigm, named entity recognition (NER) enables models to label tokens as part of entity spans and identify key entities within text (Eisenstein, 2019, p. 175). By isolating key actors, organizations, or geopolitical references, models can then compare extracted entities against trusted knowledge bases for verification (Spalenza et al., 2021). Several studies demonstrate the effectiveness of this approach for misinformation detection. Al-Ash and Wibowo (2018) achieved 96.74% accuracy in distinguishing authentic from fake articles in Indonesian-language news by incorporating term frequency, phrase detection, and NER (Al-Ash & Wibowo, 2018).

In contrast, other research suggests that NER may not always be the most decisive factor for model performance. González Silot et al. (2025) built on this method by using SHAP-based explainability, a technique that identifies which features most influence a model's predictions, to

pinpoint misleading signals in classification tasks. They demonstrated that removing non-informative elements and replacing named entities with placeholders improves generalization and boosts performance across external datasets (González Silot et al., 2025)

The inconsistency in NER effectiveness highlights a broader concern relevant to this study's experiments because detection features that prove effective on one content type may not transfer reliably to another. NGC, aligned with the transmission model of communication, often contains formal entity references (official names, organizational titles, geopolitical entities) for the sake of clarity, which NER systems are trained to recognize. UGC, by contrast, may reference entities informally, through nicknames, hashtags, or community-specific terminology, reducing the effectiveness of NER-based approaches. This suggests that models relying heavily on NER may perform differently across UGC and NGC, a hypothesis this study's approach is uniquely positioned to test.

In a fact verification task, models must propose how things are in the real world (Eisenstein, 2019, p. 397). To achieve this, models may incorporate retrieval-augmented generation (RAG) (Lewis et al., 2020), which enables them to reference external information sources rather than relying solely on their training corpora. Fact verification and misinformation detection are closely related tasks, but their methodologies differ. For instance, fake news detection typically frames the problem as text classification, whereas fact verification relies on information extraction (Li & Zhou, 2020).

Research demonstrates the potential effectiveness of combining the two approaches to combat misinformation. Li and Zhou (2020) use a summarization model to condense lengthy news articles into short claims, which are subsequently evaluated by a fact verification model

trained on large-scale datasets. This approach reduces reliance on large, labeled datasets while still achieving competitive performance on fake news detection benchmarks. Similarly, Ahmed (2021) combines text classification with fact-checking by automatically identifying "check-worthy" statements within political debates and news articles. By combining classification with contextual features and external sources such as Wikipedia, Ahmed (2021) demonstrated that prioritizing claims for verification improves detection accuracy and creates a practical bridge between automated detection and fact-checking workflows.

These combined approaches demonstrate the significance of evaluating detection performance across content types. Fact verification pipelines assume claims are sufficiently structured to be extracted, summarized, and matched against external sources. This assumption holds more reliably for NGC, where claims are often presented as declarative statements aligned with journalistic conventions. UGC, however, may embed claims within conversational threads, sarcastic commentary, or fragmented posts that resist straightforward extraction. If fact-verification techniques are central to a detection model's architecture, then performance differences between UGC and NGC may reflect not the model's capability but rather the differential applicability of the underlying verification pipeline.

Advancements in transfer learning, an approach that adapts pre-trained encoders to specific platforms and languages, have produced significant performance gains in misinformation detection. Using this approach, Reshi and Ali (2025) demonstrate that transfer learning enables pre-trained models to adapt effectively to platform-specific data, outperforming leading models on the Fakeddit (Nakamura et al., 2019) benchmark dataset. Similarly, Özçelik et al. (2023) evaluated cross-lingual transfer. They demonstrated that multilingual pre-trained

transformers can be adapted for misinformation detection in low-resource languages when the data contexts of both languages are relatively similar. Their findings also reveal that the data domain has a significant impact on model performance. They show that the contextual similarity between datasets is more important than language family or resource availability: transfer between Arabic and Chinese in the same domain outperformed transfer between English and Polish (same language family) across different domains.

Using domain-specific fine-tuning, Kim et al. (2022) observe significant gains in accuracy on their BERTweet-large model trained on a garlic/COVID-19 subset, compared with more broadly fine-tuned rumour-detection models. Qin et al. (2024) took this approach further by introducing adversarial and regularization strategies to fine-tuning. Adversarial strategies involve adding small changes to model inputs during training to make the model more robust to noise or misleading patterns. In contrast, regularization strategies impose constraints to prevent overfitting to the training data. Using these approaches, they observed reduced overfitting and improved cross-domain performance (Qin et al., 2024).

Together, these findings demonstrate the range of successful approaches to automated misinformation detection. However, they also consistently highlight that, regardless of the underlying framework, model performance remains highly dependent on the quality, scope, and representativeness of the available training data. This dependency is central to the rationale of the present study. If detection models are predominantly trained and evaluated on datasets that do not distinguish between UGC and NGC, their reported performance metrics may obscure meaningful variation in how well they generalize across content types with different structural and communicative characteristics. By evaluating models separately on UGC and NGC, this

thesis addresses this gap. It provides insight into whether current detection approaches are equally effective across the diverse content encountered in real-world social media environments.

### *2.6.3 Structural and contextual limitations of NLP in UGC environments*

LLM-based NLP systems falter when evaluated on UGC extracted from environments such as Twitter (Liu et al., 2024). The content on these platforms is characterized by brevity, informality, and high context-dependency. These features are relatively foreign to the structured datasets on which most models are trained. Bag-of-words approaches, for instance, treat words as independent features, stripping them of contextual function (Eisenstein, 2019, p. 67). This makes them a poor fit for detecting embedded misinformation in 240-character bursts laced with sarcasm, metaphors, and domain-specific slang.

Compounding these issues is the inadequate representation of diverse communicative registers in training corpora. News datasets, political speeches, and Wikipedia articles dominate LLM training pipelines (Alnabhan & Branco, 2024), leaving a representational vacuum around the types of community-embedded discourse where misinformation often proliferates. This imbalance has direct implications for the present study. If pre-training corpora overrepresent communication that reflect the values underpinning the transmission model of communication, models may develop linguistic representations that are better suited to detecting misinformation in NGC than in UGC. In contrast, UGC may present linguistic patterns that models have encountered less frequently during training and are therefore less equipped to classify accurately.

The limitations outlined above suggest that LLM-based information detection cannot be meaningfully understood through a purely technical lens. Instead, their performance must be

situated within a broader sociotechnical framework that accounts for the conditions under which information is produced, disseminated, and interpreted. The technical literature on LLMs and NLP reports considerable progress in text classification and fact verification. However, these models underperform when exposed to informal, ambiguous, or short-form discourse, particularly in user-generated environments.

For example, Liu et al. (2024) introduce a RAG framework for cross-domain misinformation detection that leverages in-context learning based on affective information. They observe that LLMs using their RAEmoLLM framework outperformed fine-tuned small language models (SLMs) on news-based datasets but underperformed on a dataset that consists of user-tweets labeled as rumors or non-rumours and exhibited mixed performances on a soft classification conspiracy dataset. Such performance discrepancies prompt our motivation to critically evaluate whether content type and domain-specific characteristics exert a measurable influence on LM classification performance. By constructing parallel datasets that preserve this distinction and evaluating models across both content types and multiple domains, this research provides empirical evidence of the extent to which current detection systems are equipped to handle the structural and communicative diversity of real-world social media environments.

## 2.7 AI-based misinformation detection approaches

This section shifts the focus from LLM architectures and NLP capabilities (as outlined in Section 2.5) to the applied methods used in misinformation detection research. It examines key learning paradigms, model tuning strategies, and prompt-based learning approaches. This section also pays particular attention to their applicability and limitations in multi-domain and user-generated content contexts. Advancements in AI (Vaswani et al., 2017), coupled with an

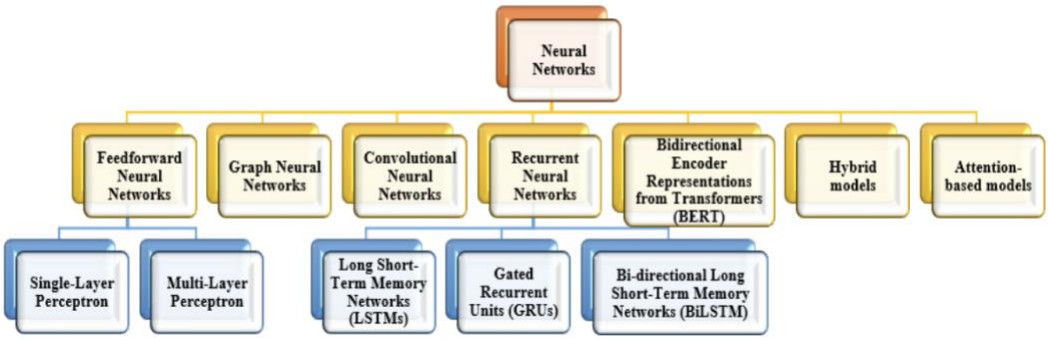
increasingly volatile sociopolitical climate, have driven substantial interest in AI applications for misinformation detection. The literature reviewed in this section highlights ML and DL techniques for scalable, automated and adaptive misinformation detection. It also explores persistent challenges inherent to this task, including dataset scarcity, model generalizability issues, and the high cost of developing new models from scratch.

### *2.7.1 Learning paradigms*

Approaches to misinformation detection can be grouped into two main categories: ML and DL. ML encompasses both supervised and unsupervised paradigms, which are explained in further depth in the following paragraphs. DL, a subset of ML, relies on neural architectures with many hidden layers. These architectures include CNNs, RNNs, and, more recently, transformer-based models. At the foundation of these architectures is the neural network, a ML model that stacks interconnected units (or nodes) in layers and learns patterns from data. In a neural network, weights that control how strongly an input feature influences a decision and biases are built-in values that skew the decision threshold itself (IBM, 2025). In misinformation detection on social media, for example, posts with certain words can make a model more likely to classify a post as false.

**Figure 2**

*Taxonomy of the main NN categories used for FND*




*Note.* 1 By Mohammed Q.M. Alnabhan [Image]. Advancing Cross-Domain Fake News Detection: Enhanced Models to Improve Generalization and Tackle the Class Imbalance Problem. University of Ottawa, March 12, 2025, https://ruor.uottawa.ca/server/api/core/bitstreams/2d0

DL architectures have become the dominant approach to misinformation detection due to their superior accuracy and robustness in handling complex linguistic patterns (Mouratidis et al., 2025). In a systematic comparison between traditional ML classifiers and DL models, Mouratidis et al. (2025) found that transformer-based BERT models consistently outperformed other methods in complex misinformation scenarios. Their findings help demonstrate why DL, particularly transformer models, are so widely adopted in research today. Pre-trained transformer-based models also provide convenient starting points for researchers who lack the computational or financial resources to develop their own models from scratch. For example, Alquadi et al. (2025) proposed a multi-stage transfer learning framework that leverages the

strengths of pre-trained LLMs tailored for binary classification of misinformation. Using this approach, they found significant improvements in detection accuracy (3.9%) in limited data scenarios.

Binary classification is a prominent supervised learning task in which models are trained to distinguish true from false claims using annotated inputs. Supervised approaches to misinformation detection have been explored using a variety of different features. In 2011, Castillo et al. analyzed Twitter (Now X) by classifying content from pre-selected bursts of newsworthy topics over two days. Using a supervised approach, they manually differentiated news content from conversational discourse, including personal opinions and user replies. By incorporating features derived from user behaviour, their study demonstrated the capacity to accurately distinguish between the propagation patterns of credible and non-credible information within the platform. In 2017, Wang presented LIAR, a dataset for fake news detection (FND) comprising 12.8K manually labeled short statements from PolitiFact. Using this dataset, the author designed a hybrid CNN to integrate metadata into text and achieved significant improvements in detection accuracy (Wang, 2017). This marked a significant contribution to FND research, and despite being published before the widespread adoption of transformer-based models, it is still commonly used as a benchmark in 2025. Despite supervised learning's effectiveness and task accuracy, the reliance on large pre-labeled datasets makes this approach very labour-intensive (Yang et al., 2019). As a result, misinformation detection research has encountered a significant bottleneck due to the scarcity of labeled datasets (Shu et al., 2017). This issue is complicated further by the weaknesses within many datasets in the field, which are

characterized by significant class imbalances that embed biases within classification models (Alnabhan & Branco, 2024) (see Section 2.8.2).

Unsupervised methods aim to reduce the time-consuming process of manually labeling datasets by using other features within text. For example, Yin et al. (2023) proposed an approach that uses both the context and content of news propagation as self-supervised signals. Using these signals, they found increases in accuracy on the PolitiFact (3.24%) and GossipCop (18.81%) datasets compared to other leading performances. In contrast, Shin et al. (2023) clustered news articles by topics related to the Ukrainian-Russian war, hypothesizing that topics with greater variance were more likely to contain fake news. Their findings revealed that unsupervised content clustering is insufficient for effectively detecting fake news with BERT-based models, demonstrating the challenges of entirely foregoing textual features (Shin et al., 2023).

To address the limitations of both supervised and unsupervised techniques, some researchers have turned to hybrid approaches for automated misinformation detection. Benamira et al. (2019) proposed a graph-based semi-supervised model that combines word embeddings with graph neural networks. Their framework captures latent representations of news articles using word embeddings and a graph-based representation scheme to model contextual similarities among articles. Results showed that supervised methods outperformed traditional classifiers, particularly when trained on a limited number of labeled articles (Benamira et al., 2019). Similarly, Shaeri and Katanforoush (2023) developed a self-learning semi-supervised method that integrates transfer learning with sentiment analysis. Their study reported significant

accuracy gains compared to both traditional ML methods and DL models without transfer learning or sentiment features. However, because the goal was to demonstrate the effectiveness of transfer learning for sentiment analysis, the authors first had to source a baseline dataset. They described this process as particularly challenging, noting issues with inactive accounts and deleted content, which complicated the development of an equivalent dataset with sentiment encoding (Shaeri & Katanforoush, 2023). Overall, while hybrid approaches reduce some of the laborious aspects of fully supervised methods, they remain constrained by dataset scarcity.

The tuning methods used in automated misinformation detection research represent another distinction amongst the approaches used in this field. Fine-tuning is a common machine-learning technique that makes small parameter adjustments to pre-trained LLMs to improve their performance on specific tasks (Qin & Zhang, 2024). Kim et al. (2022) demonstrated its effectiveness by fine-tuning five BERT-based models on two datasets: a general COVID-19 rumour dataset and a garlic-specific misinformation dataset. They then evaluated the models on tweets mentioning both garlic and COVID-19. While all models fine-tuned on the general COVID-19 dataset performed poorly (maximum accuracy of 64.7%), those fine-tuned on the garlic dataset achieved significant gains, reaching 91.1% accuracy. The stark contrast in performance between datasets used for tuning reveals the impact of appropriately curated data.

Qin and Zhang (2024) examined fine-tuning challenges in greater depth, focusing on strategies to reduce overfitting, a phenomenon in ML classification where fine-tuned models overly rely on their training set and struggle to generalize to unseen data (Qin & Zhang, 2024). They proposed an adversarial fine-tuning strategy based on feature regularization, which reduces

overfitting by encouraging weights to become smaller, thereby reducing their influence on model decisions. Their results showed improved generalization, with performance increases rising from 61% to 73% for fine-tuned BERT-based models (Qin & Zhang, 2024). However, by comparing unmodified LLMs with their fine-tuned variants, Yang et al. (2024) demonstrated that outcomes depend heavily on the strategy and task. They reported improved generalization when in-context learning was incorporated, with fine-tuned models outperforming the unmodified baseline on in-domain tasks. However, their results indicate that fine-tuned LLMs do not consistently improve with in-context learning, and in some cases, models without it achieve higher performance. Moreover, Yang et al. (2024) reported that, on out-of-domain datasets, fine-tuned models underperformed baseline variants in generalization tasks but achieved superior results in classification tasks. These findings suggest that fine-tuning is not universally optimal; its effectiveness depends on the data, task type, and tuning strategies used.

Instruction tuning provides a alternative approach to traditional supervised fine-tuning. Instead of adjusting parameters solely based on labeled input-output pairs, models are trained on datasets that pair natural-language instructions with appropriate responses (Zhang et al., 2023). This method has the advantage of being computationally efficient and allowing for greater control and predictability compared to baseline LLMs (Zhang et al., 2023). Instruction tuning is distinct from Reinforcement Learning from Human Feedback (RLHF), which uses human preference rankings to optimize model outputs through reinforcement learning. While RLHF shapes models to align with human values and preferences, instruction tuning teaches models to follow task descriptions expressed in natural language.

Although instruction tuning has achieved strong results in classification tasks, it has garnered considerable skepticism from researchers regarding what models actually learn from instructions. Kung and Peng (2023) conducted an empirical study of instruction tuning to explore whether LMs truly capture the semantic content of instructions. They found that models trained with simplified task definitions achieve comparable performance to those trained with robust examples and detailed instructions. These findings suggest that the performance gains observed with instruction tuning may stem from models picking up on superficial patterns, such as the output format, rather than genuinely understanding the task instructions (Kung & Peng, 2023). This raises questions about the mechanisms underlying instruction tuning's effectiveness and whether models are learning to follow instructions or simply pattern-matching at the surface level.

### *2.7.2 Prompting strategies*

In AI, a prompt is natural-language input provided to a model to elicit a specific output (Google Cloud, 2026). The quality and format of a prompt directly influence model performance, and well-crafted prompts can even override constraints embedded within chat-based LLMs. For instance, Liu et al. (2024) demonstrated that carefully engineered prompts can bypass ChatGPT's safety restrictions across 40 use-case scenarios. Similar effects have been documented in misinformation detection, where researchers find that prompt formulation can significantly shape model outputs and overall detection performance (Hu et al., 2023; Cao et al., 2025).

Beyond the influence of prompt engineering on safety constraints, research also examines whether LLMs genuinely comprehend the semantic content of prompts or match surface

patterns. Jang et al. (2022) explored this question using negated prompts, evaluating LM outputs when input prompts were modified to convey the inverse meaning. For example, by swapping "complete the sentence with the appropriate ending" to "complete the sentence with the inappropriate ending," they found that all LM types performed worse on negated prompts and showed a significant performance gap compared to human performance (Jang et al., 2022). These findings indicate that models may rely on shallow pattern recognition rather than deep semantic understanding, suggesting that there is considerable work to be done in understanding how and why specific prompt formulations elicit performance gains.

The prompts used in the vast majority of misinformation detection research can be summarized in three major categories. Zero-shot prompting refers to a setting where the model is given a task description without examples. In a zero-shot setting, researchers often use role-playing. Shanahan et al. (2023) contend that role-play is central to understanding chatbot LLMs, while Liu et al. (2024) demonstrate that it can reveal weak spots in their restriction parameters. Few-Shot Prompting (Brown et al., 2020) is a paradigm in which the prompter provides the model with task-specific instructions and several labeled examples. Brown et al. (2020) found that scaling up LMs through a series of task-specific examples greatly improved GPT-3's performance on a wide range of NLP datasets compared to a zero-shot setting (Brown et al., 2020). Both zero-shot and few-shot prompting can be expanded upon by adding a chain-of-thought (CoT) element. In this approach, the prompter encourages the model to reason sequentially and out loud. A common approach is to add an eliciting sentence to the prompt instructions, such as "Why don't we approach this task step by step?" (Hu et al., 2023). Wei et al. (2022) found that this approach significantly improves the ability of LLMs to perform

complex reasoning tasks. Similarly, Cao et al. (2025) found that a domain-specific CoT prompt significantly enhances the reasoning capabilities of LLMs, improving misinformation detection on finance content.

### *2.7.3 Model selection considerations*

In FND research, model selection varies significantly across studies. However, the introduction of transformer-based models, such as BERT (Devlin et al., 2019), led to their widespread adoption in FND research, as they outperform traditional ML techniques and DL models on these tasks. As newer models with improved computational efficiency and larger parameter sizes are released, researchers have begun exploring whether scaling up improves detection performance. In this context, comparisons have been made between SLMs and LLMs. It is important to note that there is no clear consensus regarding the size boundaries between SLMs and LLMs. This is partly due to the rapid pace at which larger models are being developed, which continually shifts the definitions of "small" and "large." For instance, upon its release, BERT (Devlin et al., 2019) was widely considered a large model at 108 million parameters. However, GPT-3 (Brown et al., 2020), with 175 billion parameters, makes BERT appear relatively small, leading to models in this range sometimes being referred to as SLMs.

The large parameter count of LLMs makes them difficult to fine-tune effectively for specific tasks (Hu et al., 2024). As a result, SLMs are often selected for misinformation detection because of their manageable size, computational efficiency, and performance. Sheng et al. (2022), for example, introduced the News Environment Perception (NEP) framework and evaluated it using traditional classifiers and BERT-emo (Zhang et al., 2021). This pre-trained

variant combines emotional features with BERT representations for classification. Their framework incorporated signals from the broader news environment, such as macro-level topic popularity and micro-level novelty. They found that the base pre-trained BERT-emo outperformed traditional classifiers on Chinese datasets, both with and without NEP, and achieved the best detection performance on English datasets when paired with NEP (Sheng et al., 2022). Similarly, Kaliyar et al. (2021) proposed FakeBERT, a BERT-based deep learning approach that combines different kernel sizes and filters with the BERT to help the model handle ambiguity. In this setting, a kernel is a fixed-size matrix applied to text embeddings to identify local patterns within the input. At the same time, filters are collections of kernels that enable the model to extract multiple types of features simultaneously (Hofmann et al., 2008). Using this approach, they achieved 98.90% accuracy on the real-world FakeNewsNet dataset (Shu et al., 2020; Kaliyar et al., 2021). Together, these findings highlight the effectiveness of pre-trained variants as both efficient baselines and adaptable components within new detection frameworks.

Although SLMs' smaller size offers practical advantages for researchers, their limited knowledge base and computational capacity can restrict further enhancement (Hu et al., 2024). Sheng et al. (2021) observed bottlenecks in BERT-based detection when applied to misinformation on unfamiliar topics. Because BERT was pre-trained on text corpora of its time, such as Wikipedia and Google's BooksCorpus (Devlin et al., 2019), it struggled to classify news items that required knowledge beyond those sources. These constraints make LLMs a potential solution for bridging the knowledge representation gaps inherent to SLMs. However, findings on LLMs' performances make their applicability to FND unclear.

Pelrine et al. (2023) sought to clarify the role of LLMs in mitigating misinformation by evaluating their performance across multiple detection tasks. Their findings indicated that both optimized and zero-shot implementations of GPT-4 outperformed other LMs in a hard binary classification task. However, their study also reported that DeBERTa (He et al., 2020) exceeded GPT-4 in six-way classification, while SqueezeBERT (Iandola et al., 2020) surpassed it in soft classification. Notably, even in binary classification, no model achieved an accuracy or F1-score above 68.2%, and the highest score for six-way classification was 29.9%, highlighting the overall limitations of current models. Caramancion (2023) reported comparable trends while assessing five LLMs (ChatGPT 3.5, ChatGPT 4.0, Bing AI, and Bard) on a news verification task. This study reported an average detection accuracy of 65.25%, with GPT-4.0 achieving the highest performance at 71%. These findings led Caramancion (2023) to suggest that more recent models hold a relative advantage in distinguishing fact from deception.

Hu et al. (2024) reported findings consistent with prior research, noting that SLMs outperform LLMs on a comparable task. Nevertheless, the authors wished to clarify further the role of LLMs within the broader landscape of FND. Their study demonstrated that LLMs could enhance SLM performance by providing informative rationales and that combining perspective-specific prompts with a rule-based ensemble of judgments could improve LLM accuracy. Findings by Hu et al. (2024) highlight the influence of prompt engineering on LLM performance and point to the potential value of their extensive knowledge representations. Cheung and Lam (2023) investigated the use of external knowledge to improve detection and found that combining an instruction-tuned LLaMA with external knowledge retrieval enhances performance compared to instruction-tuning alone. Most notably, when evaluating LLaMA on

LIAR (Wang, 2017) and RAWFC, it outperformed traditional classification methods in both conditions, including a BERT-based model, SBERT-FC (Kotonya & Toni, 2022). These findings further demonstrate that the role of LLMs in mitigating misinformation remains insufficiently understood, indicating a need for continued research to address existing gaps in the literature.

## 2.8 Critical gaps and future directions

Despite significant advances in automated misinformation detection, challenges continue to limit the real-world applicability of existing systems. Most detection models are trained and evaluated within single domains, failing to generalize when confronted with content from unfamiliar contexts or sources. This limitation is compounded by the scarcity of diverse, publicly available datasets and the persistent quality issues within those that do exist, including class imbalance and outdated content. This section examines these challenges to identify the path forward for misinformation detection research and to situate the contributions of this study within the broader landscape of information science and communication research.

### *2.8.1 Multi-domain detection challenges*

Many existing automated methods for misinformation detection are designed for a single domain (Li et al., 2021). Researchers have found that many models experience a significant drop in detection performance when previously unseen domain content is introduced (Silva et al., 2021). To overcome these limitations, Wang et al. (2018) proposed an early detection method, the Event Adversarial Neural Network (EANN), which removed event-specific features within events and retained common features across topics on social media to improve generalization. Despite successful initial results, further investigation revealed that the EANN framework did

not actually improve generalization; instead, it adapted the learning model from one domain to another. Shu et al. (2022) used a context-aware adversarial approach for multi-modal misinformation detection. In this approach, they used unsupervised feature representations for domain adaptation, thereby allowing a domain-specific classifier to detect domain differences and to maximize domain-independent features. A subsequent expansion on this framework by Mosallanezhad et al. (2022) proposed integrating auxiliary information into a reinforcement learning-based approach, allowing the model to extract domain-independent features while concealing domain-specific ones. This approach led to better adaptation to new domains on several DL models.

Li et al. (2021) achieved 5.2% accuracy compared to baseline performance using their multi-source domain adaptation with weak supervision framework, an approach that transfers knowledge from multiple source domains through adversarial training to learn domain-invariant features. Their approach also integrates weak supervision by using researchers' knowledge to identify misinformation in the target domain using established heuristic rules. However, further analysis revealed that the effectiveness of this approach depends heavily on the similarity between domains and the quality of available data. Shi et al. (2023) sought to address this constraint through a context-aware adversarial approach that effectively handles cross-domain sample uncertainty by modelling its structural dependencies. Their findings indicate that this strategy improves domain representations and reduces the transfer of irrelevant features.

Although each of these studies marked an advancement in cross-domain misinformation detection, none evaluated state-of-the-art LLMs on their framework. Currently, their use for cross-domain misinformation detection remains underexplored. Alghamdi et al. (2024) evaluated

GPT-2 using a prompt-based approach for cross-domain evaluation and found significant improvements over baselines in zero-shot cross-domain classification tasks. However, GPT-2 does not necessarily represent the potential capabilities of newer LLMs.

Liu et al. (2024) presented one of the few studies to evaluate and compare frontier LLMs with smaller leading models such as BERT and RoBERTa on a cross-domain misinformation detection classification task. Their RAEmoLLM framework, which uses in-context learning based on affective information, was evaluated on three benchmark datasets. Using their framework, they reported a 39% improvement in few-shot settings over zero-shot results, highlighting the potential of LLMs when well-crafted context-aware prompting strategies are used. More specifically, they reported LLMs outperforming fine-tuned SLMs on AMTCele (Liu et al., 2024). This dataset consists of sections of articles obtained from various news sites and from celebrity magazines. However, the same models underperformed compared to fine-tuned BERT and RoBERTa variants on PHEME (Kochkina et al., 2018), a dataset that consists of user-tweets labeled as rumours or non-rumours, and exhibited mixed performances on COCO (Langguth et al., 2023), a dataset of tweets that fall into 12 conspiracy theory categories. Such performance discrepancies prompt the central drive of this study, to critically evaluate whether domain-specific characteristics and content structure exert a measurable influence on LM classification performance.

Liu et al. (2024) attributed LLM underperformance on PHEME to the brevity and linguistic variability of social media texts. These findings emphasize a broader issue in misinformation detection: the scarcity of robust benchmarks for evaluating and comparing model performance on noisy UGC to professionally crafted NGC. Given that misinformation frequently

originates and spreads through such channels, the lack of representative benchmarks raises concerns about the real-world applicability of existing systems.

### *2.8.2 Dataset scarcity and resource constraints*

A persistent challenge in misinformation detection research is the scarcity of publicly available labeled datasets (Shu et al., 2017; Alnabhan & Branco, 2024). While strategies such as crowdsourcing, unsupervised, and semi-supervised learning have been explored as alternative approaches (Zhang et al., 2021), these methods often serve as workarounds rather than solutions to the core issue. Consequently, the same datasets are repeatedly used across studies, resulting in limited model diversity and restricted generalizability of findings.

As interest in FND grows, many benchmark datasets have been proposed. Datasets such as LIAR (Wang, 2017) and FakeNewsNet (Shu et al., 2020) are popular choices. However, their domain-specificity and reliance on fact-checking sites limit their applicability for detecting content as it appears on social media. Many datasets contain content sourced from social media, such as PHEME (Kochkina et al., 2018), CREDBANK (Mitra & Gilbert, 2015), Fakeddit (Nakamura et al., 2019), Twitter15 (Liu et al., 2015), Twitter16 (Ma et al., 2017), some-like-it-hoax (Tacchini et al., 2017), and TruthSeeker (Dadkah et al., 2023). However, many datasets are not publicly available to users, and some sources social media content from fact-checking sites and include a significant amount of content directly from professional news accounts. As a result, model performance on such datasets is not necessarily representative of detection performance on content posted by everyday users.

Beyond scarcity, the quality of existing misinformation datasets presents additional obstacles. Alnabhan and Branco (2024) highlight the issue of class imbalance in popular

datasets. Their review highlights that models trained on imbalanced datasets are prone to overfitting to the majority class, leading to excessive false positives. Class imbalance not only skews model predictions but also inflates accuracy scores, making the metric misleading in contexts where true claims dominate. High accuracy in such cases may mask poor recall for misinformation, rendering models ineffective at their core task (Alnabhan & Branco, 2024).

This puts into question the reliability of models in real-world scenarios where identifying false claims is the primary objective. For example, despite the PHEME dataset (Kochkina et al., 2018) standing out as one of the few benchmark datasets that consists largely of "authentic" UGC, it suffers from significant limitations. Within, non-rumour tweets nearly double the presence of rumour tweets, and health-related rumours are limited to 14 tweets about the Ebola virus that are all classified as true. Not only does this demonstrate the imbalanced nature of the dataset, but it also shows that the topics within it are outdated. Performance issues on PHEME, including those reported by Liu et al. (2024), demonstrate the critical need for deeper investigation into LLM capabilities on current UGC, as reflected in a balanced, domain-diverse dataset.

### *2.8.3 Summary and path forward*

This literature review has traced the evolution of misinformation as a sociotechnical challenge shaped by the transformation of social media platforms, the participatory nature of digital communication, and the integration of increasingly sophisticated AI systems into information ecosystems. Several key insights emerge from this review that inform the present study's approach and rationale.

This literature review helps position misinformation as a phenomenon embedded within specific communicative contexts and social practices. Carey's (2009) distinction between transmission and ritual models of communication provides a framework for understanding why content produced by news organizations (NGC) and everyday users (UGC) may differ fundamentally in structure, purpose, and linguistic features. While NGC typically aligns with transmission-oriented communication—emphasizing accuracy, authority, and informational clarity—UGC more often reflects ritual-oriented communication centred on participation, affiliation, and social identity. This distinction has profound implications for misinformation detection, suggesting that a single detection approach may not perform equally well across content types that serve different communicative functions.

The literature on information-seeking behaviour reveals that traditional models (Wilson, 1981; Ellis, 1989; Kulthau, 1998) were developed for contexts characterized by active, intentional search and critical evaluation. These assumptions no longer hold in social media environments, where users encounter information passively through feeds curated by personalized recommender systems and rarely engage in verification behaviours (Neely et al., 2021; Avram et al., 2020). The shift toward passive consumption is further complicated by the integration of LLMs into information retrieval, where AIOs and chatbot interfaces present synthesized conclusions that users increasingly accept without consulting primary sources (Pew Research Center, 2025). This transformation underscores the urgency of developing detection systems capable of handling the diverse, unverified content that circulates on social media platforms.

Moreover, our review demonstrates that advances in NLP have enabled powerful capabilities for misinformation detection, but these capabilities come with documented limitations. LLMs exhibit task-dependent performance (Spatharioti et al., 2025), sensitivity to prompt formulation (Fernández-Pichel et al., 2024), and embedded biases that reflect their training data (Fang et al., 2024). Critically, the dominant pre-training corpora create a representational imbalance that favours transmission-oriented content over the informal, context-dependent discourse characteristic of UGC. Research by Liu et al. (2024) demonstrates this imbalance, showing that LLMs outperform fine-tuned models on news-based datasets but underperform on Twitter-based rumour datasets. These findings suggest that current detection systems may not generalize reliably across the full spectrum of content encountered in real-world social media environments.

The literature on domain-specific misinformation shows that detection challenges vary substantially across topical contexts. Health misinformation presents unique challenges due to the social support dynamics that drive users to seek information on platforms despite widespread skepticism (Neely et al., 2021; Thackery et al., 2013). Political misinformation operates within heavily polarized information ecologies where partisan alignment shapes both content reception and propagation (Allcott & Gentzkow, 2017; Cinelli et al., 2020). Misinformation in conflict contexts carries distinct narrative conventions and high social stakes (Lewandowsky et al., 2013). Despite these documented domain-specific dynamics, most detection research evaluates models on single-domain datasets or aggregates performance across domains without examining within-domain variation across content types. This leaves open the question of whether detection systems perform consistently across both content types and topical domains.

Finally, research on LLM deployment strategies reveals ongoing uncertainty about optimal approaches to misinformation detection. Fine-tuning improves in-domain performance but risks overfitting and reduced generalization to out-of-domain data (Qin & Zhang, 2024; Yang et al., 2024). Instruction tuning offers efficiency gains but may rely on superficial pattern matching rather than genuine task understanding (Kung & Peng, 2023). Prompting strategies—whether zero-shot, few-shot, or CoT—significantly influence performance, yet their effectiveness varies by task complexity and domain (Brown et al., 2020; Wei et al., 2022; Cao et al., 2025). These inconsistencies highlight the need for systematic evaluation of how different deployment approaches perform across diverse content types and domains.

Despite substantial progress in misinformation detection research, critical gaps remain. Most notably, existing evaluation practices treat social media content as a homogeneous category, without distinguishing between UGC and NGC or accounting for the communicative functions these content types serve. As a result, leading benchmark social media datasets aggregate performance metrics that may obscure meaningful variation in how well models generalize across content with different structural and linguistic characteristics. Furthermore, while domain-specific detection challenges are well documented, research examining whether content-type performance differences persist across multiple domains remains unexplored. This creates uncertainty about the real-world applicability of detection systems intended for deployment on platforms where UGC and NGC coexist and where misinformation spans health, politics, conflict, and other high-stakes domains.

The present study addresses these gaps through a targeted evaluation designed to answer whether and how detection performance varies across content types and domains. Specifically,

we evaluate LLM-based detection systems and fine-tuned BERT models on parallel datasets that distinguish between UGC and NGC across health, politics, and war domains. By preserving this distinction throughout the evaluation, the study enables direct comparison of model effectiveness across content aligned with transmission versus ritual communicative functions (Carey, 2009). The evaluation further examines how prompting strategies (zero-shot versus few-shot) influence performance differences across content types and compares LLM patterns with those exhibited by supervised BERT-based baselines.

By situating detection performance within the communicative and domain-specific contexts reviewed in this chapter, this research contributes to a more nuanced understanding of automated detection systems' capabilities and limitations. It moves beyond aggregate performance metrics to examine whether detection systems reliably generalize across the structural, linguistic, and functional diversity present on social media platforms. In doing so, the study highlights the importance of designing detection approaches that maintain consistent performance across heterogeneous communication environments.

# 3. Methodology

This chapter outlines the methodological approach used to evaluate LM performance on a misinformation detection task across content types reflecting the values of different models of communication within multiple domains. Section 3.1 describes the research design and the epistemological foundations of the binary classification approach. Section 3.2 presents the dataset selection process, including the sources of claims from a mix of unlabeled datasets and fact-checking websites, and preprocessing to ensure structural consistency and proper labelling. Section 3.3 introduces the transformer architectures underlying both the BERT-based baseline models and the LLMs used in this study and clarifies the differences between encoder-only and decoder-only models. Section 3.4 describes the prompt engineering process, including the applied prompt design framework and preliminary testing to identify optimal prompt structures. Finally, Section 3.5 describes the experimental settings, computational environment, and evaluation metrics used to assess model performance.

## 3.1 Approach, design, and epistemological stance

This study employed a 2 × 3 × 2 factorial design to evaluate the detection performance of six LLMs across two content source types (UGC and NGC), three topical domains (health, politics, and war), and two prompting conditions (few-shot and zero-shot). This design enabled comparative evaluation across all combinations of content type, domain, and prompting strategy. The construction of the dataset used for these comparisons is described in Section 3.2.3.

To account for variability arising from both individual claims and model differences, results were modeled using a GLMM (see Appendix B for the complete GLMM). GLMMs

extend linear mixed models when the response variable is not normally distributed. In this case, because the dependent variable is binary (correct vs. incorrect model prediction), we used the binomial distribution with a logit link function. This approach estimates the log-odds of a successful prediction while constraining predicted probabilities to be between 0 and 1 (UCLA, n.d.).

In this study's GLMM, content type, domain, prompting condition, and their interactions were specified as fixed effects. Claim ID and model identity were included as random intercepts to adjust for claim-level differences in difficulty and baseline differences across LLMs. Including these as random intercepts improves the overall estimation of the fixed effects and their interactions.

In addition to our primary GLMM, we modeled a second GLMM to compare the performance of three encoder-only BERT-based models with decoder-only LLMs. A separate analysis was necessary because encoder-only models differ in architecture from LLMs and do not rely on prompting strategies. Therefore, BERT-based models could not be meaningfully incorporated into the primary GLMM, which included prompting strategy as a fixed effect. Instead, to address RQ4, we restricted the comparison to the zero-shot condition of the LLMs, allowing for a direct baseline comparison between encoder-only and decoder-only model families (see Section 3.3.1 for further description of transformer architectures).

The binary classification approach presupposes that each claim can be assigned a definitive truth value. To satisfy this requirement and maintain internal validity, the dataset was restricted to claims whose veracity could be independently confirmed through fact-checking.

Claims rated as 'mostly true,' 'mixed,' or otherwise ambiguous were excluded, ensuring that the true/false labels reflect clear factual distinctions rather than interpretive judgments

## 3.2 Dataset selection, management, and sampling approach

To address the gaps identified in Section 2.8, such as limited domain diversity (Liu et al., 2024), strong class imbalance, dataset scarcity (Alnabhan & Branco, 2024), and limited use of authentic UGC, this study required a dataset that covered multiple domains and included a balanced representation of UGC and NGC. To achieve this, this study applied a mixed-source approach that combined curated social media data from unlabeled datasets with claims taken from fact-checking websites (see Section 3.2.1). This section also explains how claims were curated, verified, and how labels were assigned for the binary classification task.

### *3.2.1 Selected datasets*

A survey by Alnabhan and Branco (2024) provides a current, comprehensive inventory of popular, publicly available datasets for fake news detection (FND). Their assessment helped guide this study's approach to curating and structuring data from two unlabeled datasets and fact-checking sites. Data selection was guided by two core criteria: (1) coverage of the study's target domains (Health, Politics, War) and (2) public availability. This study sought to partially address the scarcity of publicly available datasets by curating and labeling novel UGC and NGC data. To do this, unlabeled claims were sourced from the following datasets:

1. **I'm in the BlueSky Tonight (Failla & Rossetti, 2024):** This dataset totals over 235M user posts from BlueSky Social, making it a high-coverage dataset of social interactions and UGC. The authors cleaned the data by removing all personally identifiable

information (PII), manually mapped language data to the ISO 639-2 standard, and extracted server domains. These efforts meant that our only task was to curate posts with claims that could be assigned a ground-truth label corresponding to the predetermined domains.

2. **Truth Social Dataset (Gerard et al., 2023):** This dataset comprises 12 files totalling 823,927 user posts, or 'Truths', collected from February 2022 through September 2022. The authors did not anonymize usernames; thus, efforts were made to ensure that none of the sampled 'Truths' contained personally identifiable information. As a platform, Truth Social promotes free speech and conservative ideology. The dataset's content reflects these values, making it a high-coverage dataset of social interactions among the right-wing demographic.

While much of the NGC was directly sourced from these two datasets, external sampling was required to finalize and balance domain and label representation. This second approach consisted of scraping content from PolitiFact and Snopes, two popular fact checking websites. Data selection was guided by two core criteria: (1) coverage of the study's target domains (Health, Politics, War) and (2) Article headlines that contained claims that were assigned as 'True,' 'False,' or 'Pants on Fire.' To maintain the binary nature of our dataset, we clustered 'False' and 'Pants on Fire' into one 'False' case. Articles that were assigned labels such as 'Mostly True,' 'Mostly False,' or 'Mixture,' were not sampled to remain consistent with the binary nature of the classification task.

**Table 1**

*Content Source Distribution Across Sub-Datasets*

| Source dataset | Number of claims from each dataset | |
|---|---|---|
| | UGC | NGC |
| I'm in the BlueSky Tonight | 247 | 269 |
| Truth Social | 269 | 206 |
| Fact-checking site | 0 | 41 |

### *3.2.2 Dataset pre-processing*

Upon dataset selection, all entries underwent preprocessing to ensure structural consistency, class balance, factual accuracy, and reliable labeling across domains. This step was essential for producing comparable input formats for model evaluation and maximizing internal validity.

To address structural consistency, all data were given a uniform structure of five metadata fields, including post ID, claim, domain, content type (UGC or NGC), and a binary classification label (true = 1, false = 0). Data selection was restricted to English textual content with a maximum claim length of 300 characters across both UGC and NGC content. These constraints were applied to maintain internal consistency across content types and external consistency with previous studies using social media content. Specifically, older datasets sourced from Twitter were limited to 140 characters, which was raised to 280 in 2017 (Perez, 2017).

Moreover, by maintaining a consistent character count across both UGC and NGC content, claims were presented in a relatively similar manner. To address factual accuracy, due diligence was taken to ensure all unlabeled claims were assigned appropriate labels through independent fact-checking.

### *3.2.3 Dataset design*

To successfully evaluate models across content types and domains, this study manually curated and labeled 1032 claims from the sources described in Section 3.2.1. Unlike existing FND research, which often relies on pre-labeled data with significant class imbalance and limited domain coverage, this study's manual curation enabled control over both class distribution and topic representation.

The final dataset consists of two parallel sub-datasets: 516 UGC claims and 516 NGC claims. Each sub-dataset maintains a balance across the three domains (Health, Politics, and War). Within the domain, claims are evenly split between true and false labels (86 each), ensuring that models cannot achieve high accuracy simply by predicting the majority class. Beyond domain-level balance, topics within each domain were distributed as evenly as possible, with a maximum imbalance of ±4 at the topic level. This granular attention to balance strengthens the internal validity of cross-domain and cross-content type comparisons. Table 2 provides an overview of the specific topics covered within each domain.

**Table 2**

*Topic Coverage Across Domains*

| Domain | Topic |
|---|---|
| Health | COVID-19 |
| | Vaccines |
| | Drug Crisis |
| | Medicine/Public Health |
| | Sexual/Gender Health |
| Politics | January 6th (Insurrection) |
| | Immigration |
| | Elections |
| | Political Figures |
| | International Relations |
| War | Palestine/Isreal Conflict |
| | Ukraine/Russia Conflict |
| | Iranian Conflict |

Claims were curated to eliminate both cross-domain and within-domain topic leakage. Each domain was assigned a distinct set of topics, with no topic appearing in more than one

domain. Within domains, claims were selected to avoid redundancy across topics. For example, within the health domain, COVID-19-related claims were restricted to masking protocols and general information about the virus and its effects. In contrast, vaccination-related claims were treated as a separate topic, with no overlap between COVID-19 illness and vaccination content.

## 3.3 Deep learning architectures and selected models

To address gaps in earlier misinformation detection approaches, our study focused on evaluating LLMs and BERT models. As discussed in Section 2.8, there is a need for further research on LLM performance in multi-domain detection tasks, and none thus far have directly investigated their performance across two distinct forms of communicative function. Models based on the BERT architecture were selected for their demonstrated performance on these tasks and to serve as a comparison for LLMs within the task itself. The following section details the models selected for experimentation and describes the underlying architectures relevant to this study.

### *3.3.1 Transformer architecture*

The Transformer architecture was introduced in 2017 and was initially applied to a translation task (Vaswani et al., 2017). A key feature of the transformer architecture is its use of attention layers. Put simply, this means that a layer can tell the model to pay attention to specific words while ignoring others. Transformer models are composed of two blocks, the encoder and decoder. The encoder receives an input and builds feature representations, while the decoder uses these representations along with other inputs to generate a target sequence, allowing the model to produce the best possible output. Depending on the task and the model, each block can be used

independently. Through this, we can distinguish between specific types of transformer models. In our experiments, we use the following transformer models:

#### 3.3.1.1 Auto-encoding models.

Auto-encoding models exclusively use the encoder block of a transformer model. At each layer, the self-attention mechanism retains access to all the tokens in the input sequence (Raza et al., 2024). Their focus is on understanding input data by encoding its features without generating new content. As a result, they are well-suited for tasks that require understanding context, such as sentence classification tasks or FND. In this study, the following encoder-only models were fine-tuned and evaluated on our dataset: BERT-base-uncased, RoBERTa-base, and DistilBERT-base-uncased.

#### 3.3.1.2 Auto-regressive models.

Auto-regressive models use only the decoder component of a transformer architecture. Rather than retaining the complete initial input sentence, they attend to the preceding tokens in the input sequence. As a result, they are well-suited for text generation tasks, where training focuses on using preceding context to predict the next token (Raza et al., 2024). In this study, the following decoder-only models were evaluated under zero-shot and few-shot prompting conditions: Llama3-4b, Llama3-70b, Gemma3-4b, Gemma3-27b, Mistral-7b, and Mixtral8x7b.

## 3.4 Prompt engineering

Prompt engineering is the process of writing, refining, and optimizing inputs to encourage LLMs to create high-quality outputs (IBM, 2026). Unlike fine-tuned models trained on task-specific data, the LLMs in this study relied on prompting and instruction tuning to guide their outputs. This section describes the framework for designing prompts for the misinformation

detection task, the preliminary testing process to identify prompt structures, and the rationale for final prompt selection.

### *3.4.1 Prompt-engineering framework*

Given the challenges of fine-tuning large-scale models, this study adopts prompt engineering techniques to align pre-trained LLMs with the misinformation detection task in zero- and few-shot settings. Few-shot prompting, in which models are presented with labeled examples before making predictions, has been shown to improve performance without modifying model weights (Liu et al., 2024). Prompt structures were designed using the CLEAR Framework (Lo, 2023), which emphasizes concise, logical, explicit, adaptive, and reflective design principles. By avoiding content type and domain-specific terminology, this approach supports standardized evaluation across experimental conditions.

While parameter-efficient fine-tuning methods such as QLoRA can adapt LLMs with reduced computational overhead, they introduce task-specific learned parameters. As this study prioritizes consistent, weight-invariant evaluation across model architectures and training paradigms, prompt-based alignment was selected to preserve methodological comparability.

Following the CLEAR framework, multiple prompt structures were evaluated in a preliminary experiment on a 10% sample of the final dataset. This subset was manually curated to ensure balanced representation across content types and domains. This process enabled the identification of prompt formulations that performed consistently well across all evaluated models. The following prompts were selected for final experimentation due to their superior overall performance:

- Prompt 1: "Classify the following [Domain] claim from [Structure] as true or false: [Claim], respond with only one word (true/false)."
- Prompt 2: "Determine the factual accuracy of the following [Domain] claim: [Claim], respond with only one word (true/false)."

Both prompts underwent further modifications to test whether observations from previous research could yield even greater performance gains. For instance, Chen et al. (2025) observed that role-playing with LMs improved task performance. Therefore, variations such as “You are a fact-checking assistant” and “You are a misinformation detection system” were tested to see whether defining the model’s role could improve performance. However, both instances reduced F1 Scores across multiple models, suggesting that more concise prompts yielded better performance.

## 3.5 Experimental settings

The methodology consisted of two primary evaluation conditions for LLMs (zero-shot and few-shot prompting). In both conditions, LLMs were evaluated on a stratified dataset that balanced content source types (UGC and NGC) and topical domains (Health, Politics, and War).

In the zero-shot condition, models received task instructions without labeled examples. In the few-shot condition, models received the same task instructions along with four labeled examples: one true and one false instance each content source type (UGC and NGC) (see Appendix A for full prompt specifications).

Two distinct evaluation procedures were used to support the two GLMM analyses. For the primary GLMM, LLMs were evaluated in both the zero-shot and few-shot conditions using

their results from the entire dataset (n=1032). These results were used to examine the effects of content type, domain, and prompting strategy on model performance.

For the architectural comparison model (RQ4), evaluation procedures differed to ensure comparability between LLMs and encoder-based baselines. As mentioned in Section 3.1, since encoder-only models are trained using supervised learning, they cannot benefit from few-shot prompting in the same way as LLMs. Therefore, to ensure a fair comparison between model types, architecture-level analyses were restricted to LLM results in the zero-shot condition on items within the held-out test set for the supervised baseline.

More specifically, to maintain consistency with common practices in the ML literature, supervised baseline evaluations used a fixed train–test split. Prior work identifies 70:30 and 80:20 splits as the most frequently used ratios, as they provide a balance between training sufficiency and reliable evaluation (Virgazova, 2021; Tan et al., 2021). Although Sivakumar et al. (2024) report improved performance with larger training sets, the modest dataset size in this study motivated selecting a 70:30 split to preserve an adequately sized held-out test set.

Alternative evaluation strategies, such as *k*-fold cross-validation, were also considered. In *k*-fold cross-validation, the dataset is partitioned into *k* equally sized subsets, with the model trained on $k - 1$ folds and evaluated on the remaining fold. This process is repeated *k* times so that each subset serves as the test set once, and performance is averaged across all folds (Gunjal, 2021). While *k*-fold cross-validation is effective for optimizing individual models, it was not adopted in this study due to the emphasis on consistent and comparable evaluation across multiple model types. A fixed train-test split enabled controlled comparisons across architectures

and training paradigms without introducing additional computational or methodological variability.

### *3.5.1 Computational environment*

The experiments in this study were conducted using a Mac Studio Max desktop with 64 GB of LPDDR5 RAM. All experimentation was performed in Python, with Visual Studio Code (Microsoft, 2024) serving as the primary development environment. LLMs were evaluated locally using the LM Studio desktop application (Element Labs, 2026). LM Studio was selected for its support of local inference with pre-quantized LMs. The reduced memory footprint of these pre-quantized model variants enabled the evaluation of multiple LLMs within the available hardware constraints.

For the supervised baseline experiments, BERT-based models were sourced from the Hugging Face model repository and fine-tuned locally using PyTorch. Model training, inference, and evaluation relied on widely used Python libraries, including pandas, requests, sklearn.metrics, scipy, numpy, transformers, and torch. Moreover, all GLMMs were estimated in R (version 4.5.1). Model estimation was conducted using the *lme4* package (Bates et al., 2015), which fits mixed-effects models via maximum likelihood using the Laplace approximation. Moreover, the *emmeans* package (Russell et al., 2025) was used to compute estimated marginal means and pairwise contrasts, and *performance* (Lüdecke et al., 2026) and *broom.mixed* (Bolker et al., 2024) was used for model diagnostics and for extracting fixed-effect estimates where applicable. Data manipulation and visualization were conducted using the *tidyverse* (Wickham, 2023) suite of packages.

It should be noted that although the computational setup was sufficient for training and evaluating the selected models, additional resources would have enabled a wider range of model sizes. As it was, sampled models did not exceed 70B parameters due to the extensive RAM resources required for models exceeding that size. Moreover, despite their prior performance on detection tasks, we were unable to evaluate mixture-of-experts (MoE) models with more than 7 billion parameters. Therefore, it is recommended that further experimentation be conducted to evaluate LM performance in a multi-domain and content-type environment.

### *3.5.2 Evaluation metrics*

The following metrics were used to descriptively evaluate the selected models' performance on the misinformation detection task and were chosen in line with current FND research, where positive predictions refer to the model labeling the claim as true and negative predictions refer to the model labeling a claim as false.

**Table 3**

*Descriptive Metrics Used in our Experiments*

| Metrics | Formula | Description |
|---|---|---|
| Accuracy | $\frac{TP + TN}{TP + TN + FP + FN}$ | The proportion of correctly classified instances out of all predictions made. |
| Precision | $\frac{TP}{TP + FP}$ | The proportion of positive predictions that are correctly classified. |
| Recall | $\frac{TP}{TP + FN}$ | The proportion of actual positive instances that are correctly identified. |
| F1-Score | $2 \times \frac{Precision \times Recall}{Precision + Recall}$ | The harmonic mean of precision and recall. |

*Note: TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives.*

While accuracy provides an overall measure of classification correctness, it can be misleading when evaluated in isolation. Although the dataset used in this study was balanced across true and false claims between domains and content sources (see Section 3.1), accuracy does not differentiate between types of classification error. For example, a model that predicts all claims as false would achieve 50% accuracy on a balanced dataset, even though it fails to identify true claims.

For this reason, the F1-score served as the primary metric used in our descriptive performance visualizations. By incorporating both precision and recall through their harmonic mean, the F1-score penalizes models that achieve high precision at the expense of recall, or vice versa, providing a more balanced assessment of classification performance. Accuracy, precision, and recall were also retained to enable comparison with prior FND research (see Appendix C for descriptive experimental results).

### *3.5.3 Generalized linear mixed modeling*

All inferential statistical analyses reported in Chapter 4 were done using GLMMs with item-level binary correctness (correct vs. incorrect classification) as the dependent variable. This modeling approach allows for random intercepts at the item and model levels, thereby ensuring that hypothesis testing accounts for variation in difficulty across claims and baseline differences across models. Moreover, using a GLMM, hypothesis testing is performed on the full item-level data rather than aggregated summary metrics. The primary model used a binomial distribution with a logit link function:

$$\text{logit}(P(Y_{ij} = 1)) = \beta_0 + \beta_1(\text{ContentType}) + \beta_2(\text{Domain}) + \beta_3(\text{Condition}) + \beta_4(\text{Interactions}) + u_j + v_i$$

Where:

- $Y_{ij}$ represents the correctness of the prediction for claim *i* by model *j*
- $\beta_0$ represents fixed intercepts, including content type (UGC vs. NGC), domain (politics, health, war; health as reference), prompting condition (zero-shot vs few-shot; zero-shot as reference), and relevant interaction terms.

- $u_j \sim N(0, \sigma^2{}_{model})$ represents the random intercept for the model (architecture-level variability)
- $v_i \sim N(0, \sigma^2{}_{model})$ represents the random intercept for claims (item-level variability)

#### 3.5.3.1 Model architecture comparison model.

As mentioned in Sections 3.1 and 3.5, a separate GLMM was conducted to compare the performance of encoder-based BERT models with that of LLMs in zero-shot settings. As in the primary analysis, classification correctness at the claim level (correct = 1, incorrect = 0) served as the dependent variable. A binomial distribution with a logit link function was specified:

$$\text{logit}(P(Y_{ij} = 1)) = \beta_0 + \beta_1(\text{ModelFamily}) + \beta_2(\text{ContentType}) + \beta_3(\text{Domain}) + \beta_4(\text{Interactions}) + u_j + v_i$$

Where:

- $Y_{ij}$ represents the correctness of the prediction for claim *i* by model *j*
- $\beta_0$ , represents fixed intercepts, including model family (Encoder-only (BERT) vs. Decoder-only (LLM); BERT as reference), content type (UGC vs. NGC; NGC as reference), domain (politics, health, war; health as reference), prompting condition (zero-shot vs few-shot; zero-shot as reference), and relevant interaction terms.
- $u_j \sim N(0, \sigma^2{}_{model})$ represents the random intercept for the model (architecture-level variability)

- $\upsilon_i \sim N(0, \sigma^2{}_{model})$ represents the random intercept for claims (item-level variability)

# 4. Results

This chapter reports the empirical findings of this thesis. Section 4.1 presents the primary GLMM results, which examine the effects of content type, domain, and prompting condition on LLM classification performance in a binary misinformation detection task. Section 4.2 presents the second GLMM used in our study, which addresses whether content type and domain performance patterns differ between LLMs and fine-tuned BERT models (see Section 3.1 for a detailed description of the statistical models used to analyze this study's results and Section 3.5 for the experimental setup).

## 4.1 Generalized linear mixed model: LLM

This section reports the findings related to the effects of content type (RQ1), domain (RQ2), and prompting condition (RQ3) on LLM classification performance on a binary misinformation detection task. To address these questions, we modelled LLM predictions using a GLMM with a binomial distribution and a logit link, fitting the model to item-level prediction outcomes (correct vs. incorrect). Content type, domain, prompting condition, and all interaction terms were included in the full factorial model. In addition, random intercepts were included to account for model- and claim-level variability. Finally, the inclusion of interaction effects was supported by the model's improved fit relative to reduced specifications (AIC = 14437.6).

### *4.1.1 Main effect of content type (RQ1)*

A full GLMM including content type, domain, prompting condition, and their interactions significantly improved model fit relative to a reduced model excluding content type ($\chi^2(6) = 36.70$, $p < .001$). This supports the investigation of the impact of content type on LLM

classification performance by indicating that it accounts for distinct variance in classification accuracy beyond domain and prompting effects.

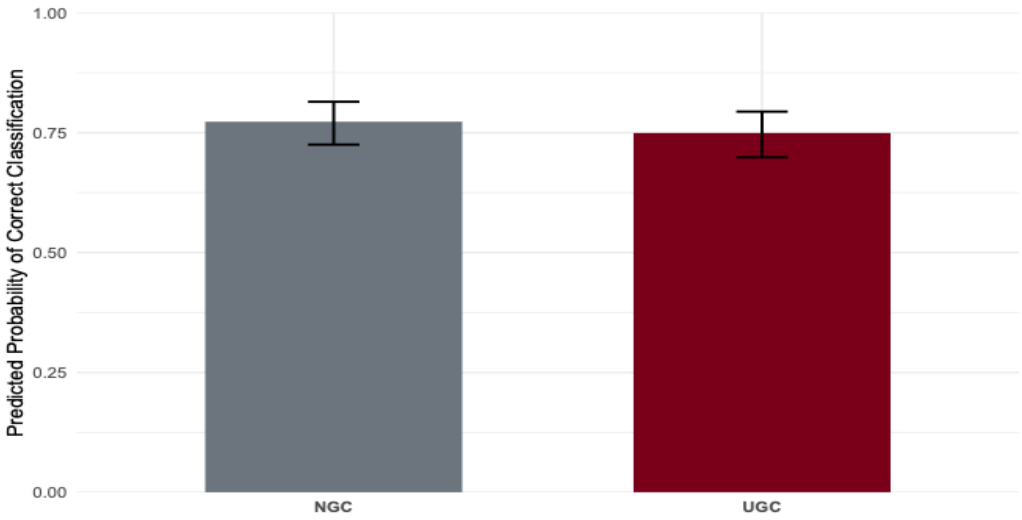
Our statistical analysis revealed a significant main effect of content type on classification accuracy ($\beta = -0.7$, SE = 0.267, $z = -2.621$, $p < .009$). Under standard treatment coding, NGC served as the reference category, meaning that the effect of content type on classification performance was measured by comparing UGC against NGC. The negative coefficient indicates that, within the reference domain (Health) and prompting condition (few-shot), models were less likely to classify UGC than NGC correctly. More precisely, under the reference conditions, the odds of LLMs correctly classifying UGC were roughly 50% lower than those for NGC, as shown by the exponentiation of the coefficient ($e^{-0.7} \approx 0.5$) (see Figure 3).

Although the reference category (in this case, NGC) serves as a comparison baseline for analysis, it does not imply prioritization, nor does it yield different overall model conclusions. A different prompting condition, content type, or domain serving as the reference category would result in equivalent conclusions being expressed relative to a different baseline.

Because higher-order interactions were present (see Section 4.1.3), this main effect should be interpreted conditionally, reflecting performance differences across the reference domain (Health) and the prompting condition (few-shot) rather than a global average effect.

**Figure 3**

*Predicted Probability of Correct Classification by an LLM Across UGC and NGC*




*Note.* 2 Bars represent model-adjusted predicted probabilities of correct classification, and bars indicate 95% confidence intervals. Predictions are averaged across domains and prompting conditions, with NGC serving as the reference category.

### *4.1.2 Content type × domain interaction (RQ2)*

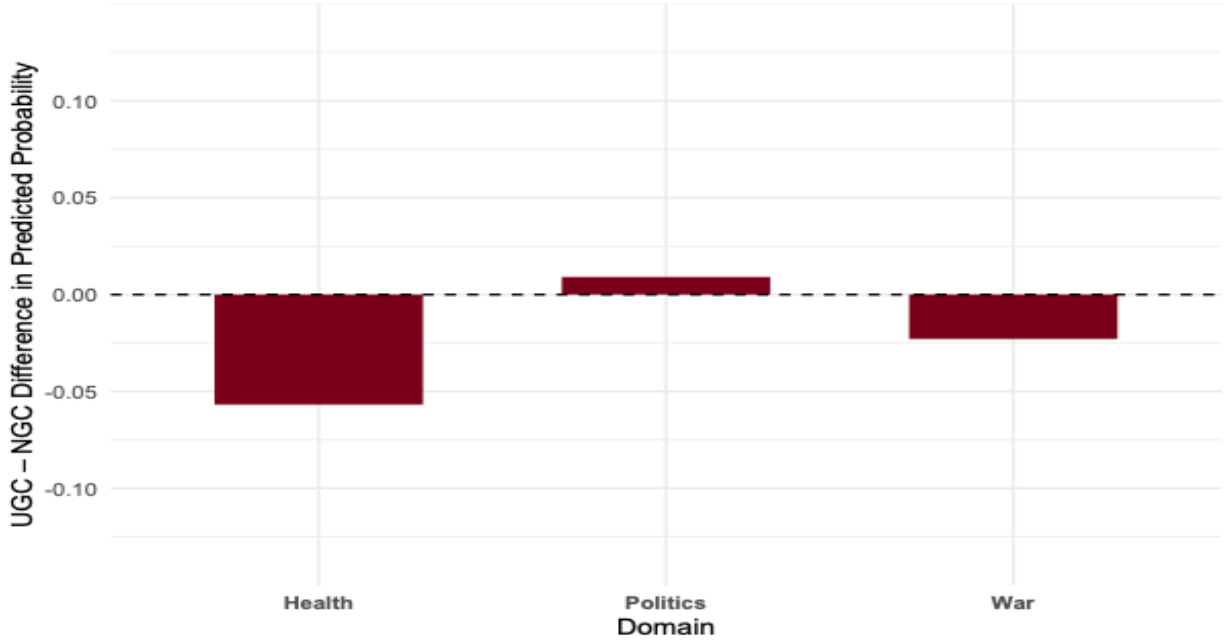
This study's second research question examined whether the effect of content type on classification performance varied across topical domains. Since Health served as the reference category in our fitted model, the interaction terms for Politics and War indicate how the NGC-UGC performance gap in those domains departs from the UGC-NGC difference observed in the Health domain.

A significant interaction was observed between content type and Politics ($\beta = 1.0084$, $SE = 0.2489$, $z = 4.052$, $p < .001$), indicating that the UGC-NGC performance gap in the Politics domain differed significantly from the corresponding gap in Health. In contrast, the interaction

between content type and War (Domain3) (β = 0.3749, SE = 0.2449, z = 1.531, p = .126) did not reach statistical significance.

**Figure 4**

*Model-Adjusted Content Type Performance Gaps by Domain (LLMs)*




*Note. 3* Negative values (those below the 0.0 baseline) indicate lower accuracy for UGC relative to NGC. Values reflect domain-specific predicted gap (not raw interaction coefficients), with other model factors held constant.

These findings indicate that the effect of content type on LLM classification accuracy differed across domains. However, as shown in the following section, this domain-specific disparity was further moderated by the prompting condition.

### *4.1.3 Three-way interaction: Content type × domain interaction × prompting (RQ3)*

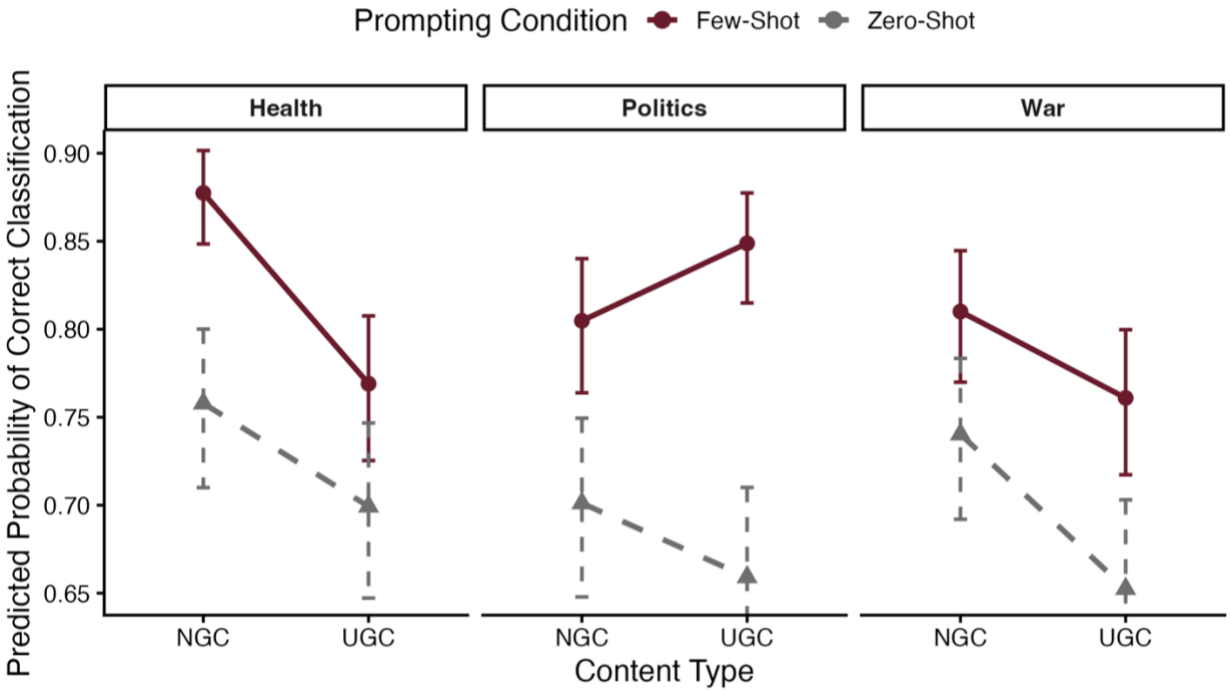
To determine whether differences across content types were jointly moderated by domain and prompting condition, we modeled a three-way interaction among content type, domain, and prompting condition. We found a significant three-way interaction ($\beta = -0.2427$, $SE = 0.1046$, $z = -2.321$, $p < .02$), suggesting that the interaction between content type and prompting condition varied across domains.

The negative coefficient ($e^{-0.2427} \approx 0.78$) indicates that the moderating effect of prompting condition on the UGC–NGC performance gap was weaker in non-reference domains relative to Health. In other words, the extent to which zero-shot prompting reduced content-type disparity in the Health domain did not generalize to Politics and War.

Given the significant three-way interaction, lower-order interactions must be interpreted conditionally within each domain. Domain-specific marginal means were examined to assess how the content type × prompting relationship varied across Health, Politics, and War (see Figure 5). Overall, our findings suggest that the magnitude of the UGC–NGC performance disparity varied across LLMs, and that its magnitude depended jointly on the topical domain and the prompting strategy.

**Figure 5**

*Content Type and Prompting Interaction Across Domains (LLMs)*




*Note.* 4 CI = 0.95; Differences in slopes across panels reflect the significant three-way interaction among content type, domain, and prompting condition.

Within the Health domain (reference category), we found a significant interaction between content type and prompting condition (β = 0.5208, SE = 0.2272, z = 2.292, p = .022). Under few-shot prompting, the predicted probability gap between NGC and UGC was 10.8 percentage points (0.877 vs. 0.769). Under zero-shot prompting, this gap narrowed to 5.9 percentage points (0.758 vs. 0.699). Exponentiating the interaction coefficient ($e^{0.5208} \approx 1.68$) indicates that the log-odds difference associated with content type was approximately 68% larger under few-shot prompting relative to zero-shot prompting within the Health domain.

Substantively, zero-shot prompting reduced the disparity between NGC and UGC, though it lowered overall classification accuracy across both content types.

In contrast, the moderation pattern differed in the Politics and War domains (see Figure 5). The attenuation effect observed in Health was weaker and did not reach statistical significance in these domains. Taken together, these domain-specific patterns explain the significant three-way interaction and demonstrate that content-type performance disparities in LLMs are shaped by the combined influence of topical domain and prompting strategy.

#### *4.1.4 Random effects*

In a GLMM, the random-intercept variance represents the extent to which outcomes vary across levels of the grouping factor (in this case, claims and models) after accounting for fixed effects (in this case, content type, domain, and prompting conditions) (UCLA, n.d.). The random intercept variance at the claim level was substantial (variance = 1.31771), indicating considerable variability in classification difficulty across claims. In contrast, variance attributable to model identity was minimal (0.09461), indicating that baseline performance differences between LLMs were relatively small after accounting for claim-level variability and fixed effects. This suggests that most of the unexplained variation in classification accuracy arises from properties of the claims themselves rather than differences between models. As such, the magnitude of item-level variance suggests that failing to account for our random effects could obscure meaningful differences in classification difficulty across various experimental settings.

## 4.2 Generalized linear mixed-model: Model architecture (RQ4)

This section presents findings relevant to RQ4, which investigate whether performance patterns across content types and domains differed between zero-shot LLMs and fine-tuned BERT-based models. To ensure comparability across model families, architecture-level analyses were conducted only within the zero-shot evaluation setting. Given that fine-tuned BERT-based models do not rely on labeled examples at test time in the same manner as few-shot prompted LLMs, we determined that restricting comparisons to the zero-shot condition was the best way to maintain methodological equivalence.

A GLMM with a binomial distribution and logit link was fitted to item-level prediction outcomes (correct vs. incorrect). A random intercept for claim ID was included to account for variability in item difficulty across claims. Initial attempts to include a random intercept for model identification resulted in a singular fit, indicating negligible variance at the model level. Therefore, the final GLMM retained only the claim-level random effect.

The full GLMM included fixed effects for content type, domain, and model architecture (LLMs vs. fine-tuned BERT), along with all corresponding interaction terms. Model fit indices (AIC = 8309.7) supported retaining the full interaction structure, thereby allowing direct tests of whether content type and domain effects were moderated by model architecture.

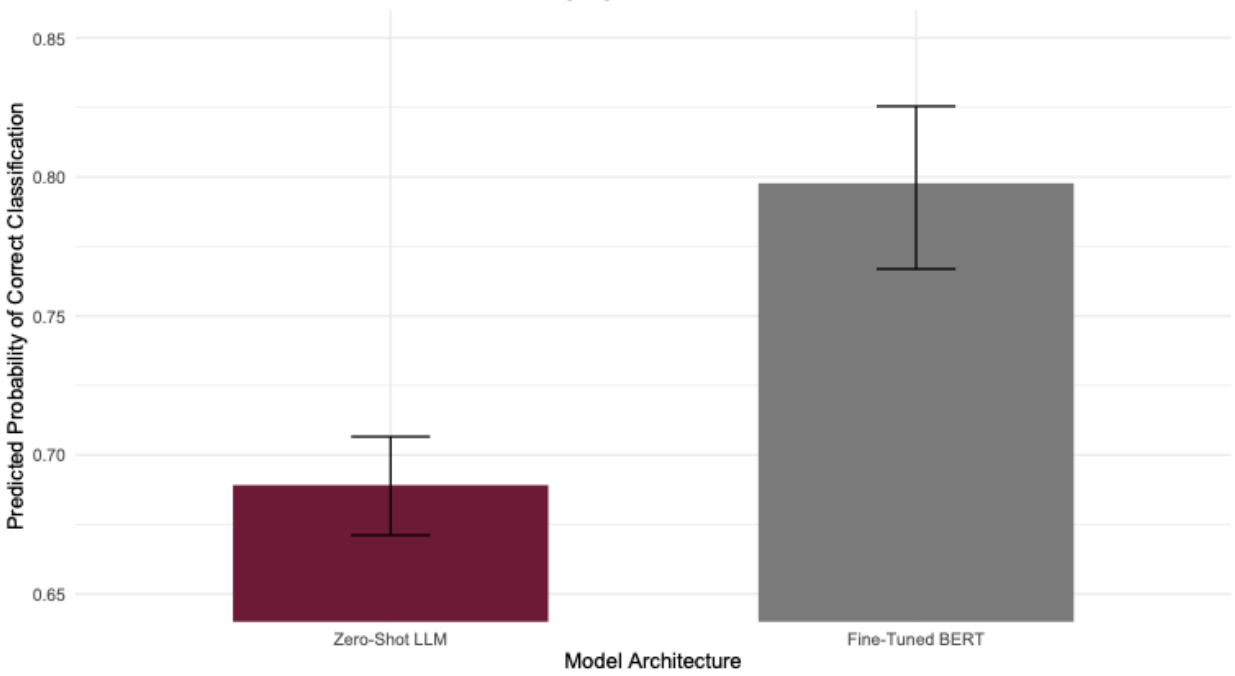
### *4.2.1 Main effect of model architecture*

Using zero-shot LLMs as the reference category (within the Health domain and NGC content type), we observed a significant main effect of model architecture on classification accuracy ($\beta = 0.7051$, SE = 0.2402, $z = 2.936$, $p = .003$). Exponentiating the coefficient ($e^{0.7051} \approx 2.022$) indicates that, within the reference condition (Health, NGC), fine-tuned BERT-based models had approximately twice the odds of correct classification relative to zero-shot LLMs. To

facilitate interpretation across conditions, we computed marginal predicted probabilities, averaged over content type and domain. Fine-tuned BERT models demonstrated higher overall predicted accuracy (0.798) relative to zero-shot LLMs (0.689), yielding an approximate 11-percentage-point advantage.

**Figure 6**

*Overall Predicted Classification Accuracy by Model Architecture*




*Note. 5* CI = 0.95

Because higher-order interactions are included in the model (see Section 4.2.4), this main effect should be interpreted as conditional in the reference domain and content type. Nonetheless, the results indicate a clear baseline architecture-level performance advantage for fine-tuned BERT models.

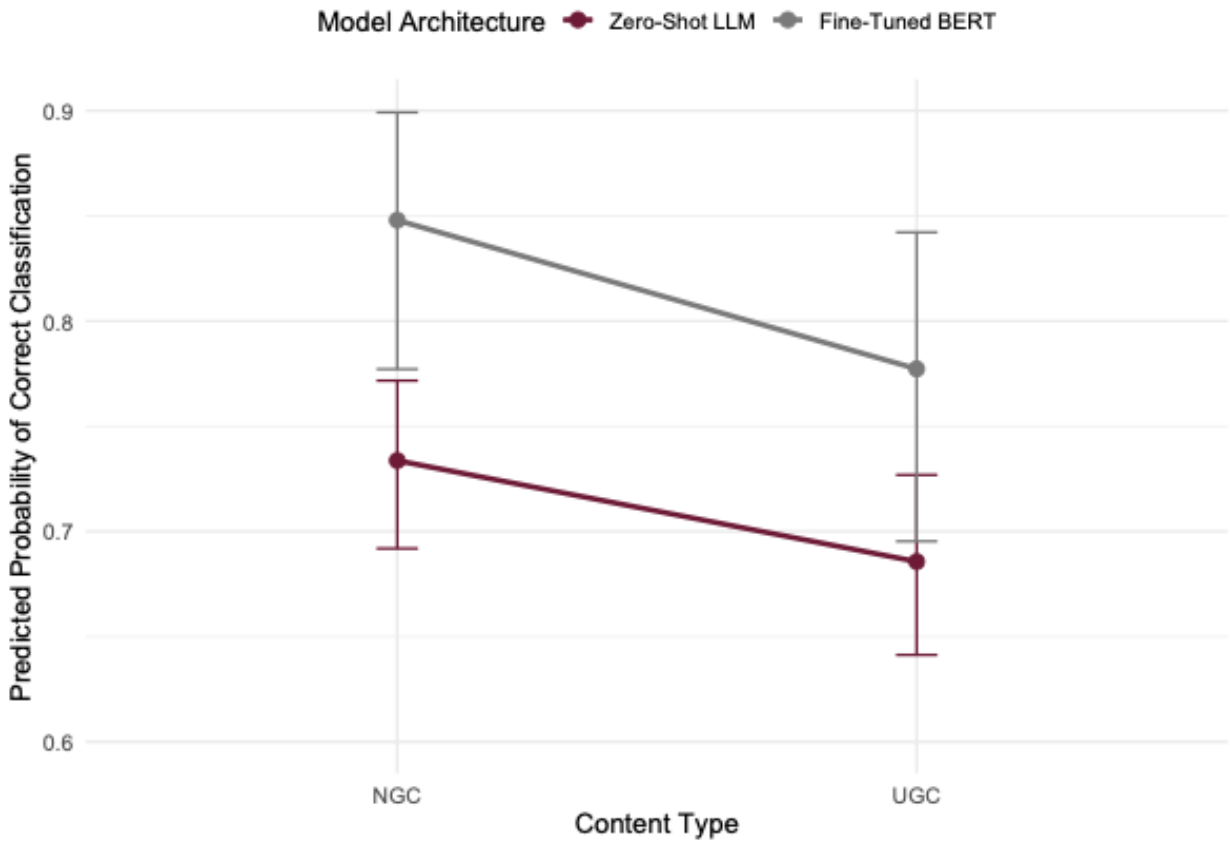
### *4.2.2 Architecture × Content type interaction*

The interaction between content type and model architecture was examined to determine whether the UGC–NGC performance gap differed across model families within the reference domain (Health). This interaction was not statistically significant (β = -0.2353, SE = 0.3262, z = -0.721, p = .471). Exponentiation of the coefficient ($e^{-0.2353} \approx 0.79$) suggests that the UGC–NGC disparity in BERT-based models was approximately 21% smaller in odds relative to the zero-shot LLMs within the Health domain, which was not considered a statistically significant deviation.

Predicted probabilities within Health indicate that both architectures exhibited a UGC-related performance reduction. Zero-shot LLM accuracy declined from 0.734 (NGC) to 0.686 (UGC), while BERT-based model accuracy declined from 0.848 (NGC) to 0.777 (UGC). Although the magnitude of the decline was slightly larger for the BERT-based models, the slopes were largely parallel (see Figure 7).

**Figure 7**

*Architecture and Content Type Interaction Within the Health Domain*




*Note.* 6 CI = 0.95

These findings suggest that although overall accuracy differed across model architectures, the magnitude of the UGC-related performance vulnerability was statistically comparable across model architectures under the reference conditions.

### *4.2.3 Domain × Architecture interaction*

The interaction between domain and model architecture was examined to determine whether domain-based performance differences varied across model families for the reference content type (NGC).

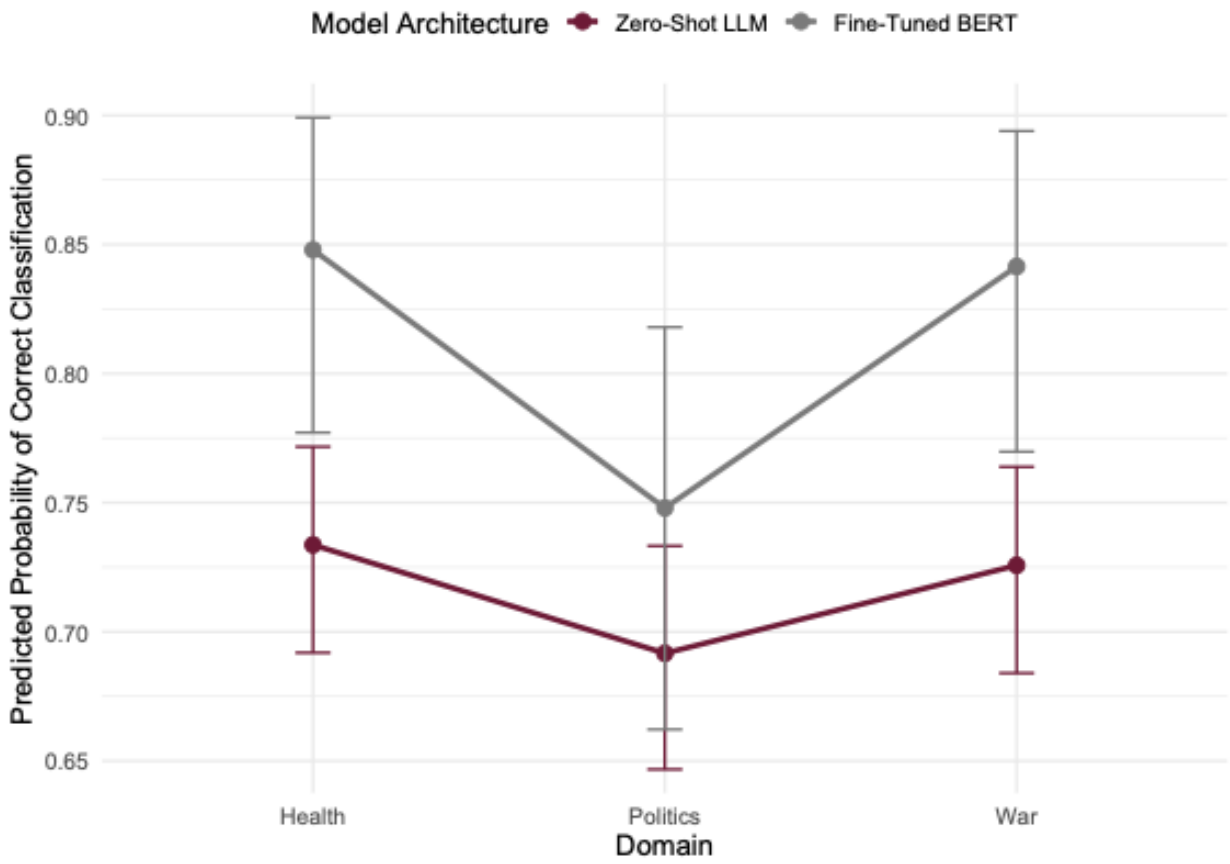
The interaction between Politics and architecture was not statistically significant (β = -0.4253, SE = 0.3196, z = -1.1331, p = .183). Exponentiation of the coefficient ($e^{-0.4253} \approx 0.65$) indicates that the Politics–Health difference in predicted classification accuracy was approximately 35% smaller in BERT-based models relative to zero-shot LLMs. However, this deviation did not reach statistical significance.

Similarly, the interaction between War and architecture on predicted classification accuracy was not statistically significant (β = −0.009, SE = 0.3364, z = −0.027, p = .979). Exponentiation of the coefficient ($e^{-0.009} \approx 0.99$) indicates virtually no deviation in domain-related performance differences between architectures.

Overall, predicted probabilities for NGC content show modest domain variation for zero-shot LLMs (Health = 0.734, Politics = 0.692, War = 0.726) and somewhat larger variation for BERT-based models (Health = 0.848, Politics = 0.748, War = 0.841). However, these differences were not statistically meaningful.

**Figure 8**

*Domain and architecture interaction within NGC*




*Note. 7* CI = 0.95

Taken together, these findings suggest that domain-based differences in classification accuracy for NGC were statistically comparable across zero-shot LLMs and fine-tuned BERT-based models.

### *4.2.4 Three-way interaction: Content type × Domain × Architecture*

The three-way interaction between content type, domain, and model architecture was examined to determine whether the domain-specific UGC–NGC performance gap differed between zero-shot LLMs and BERT-based models.

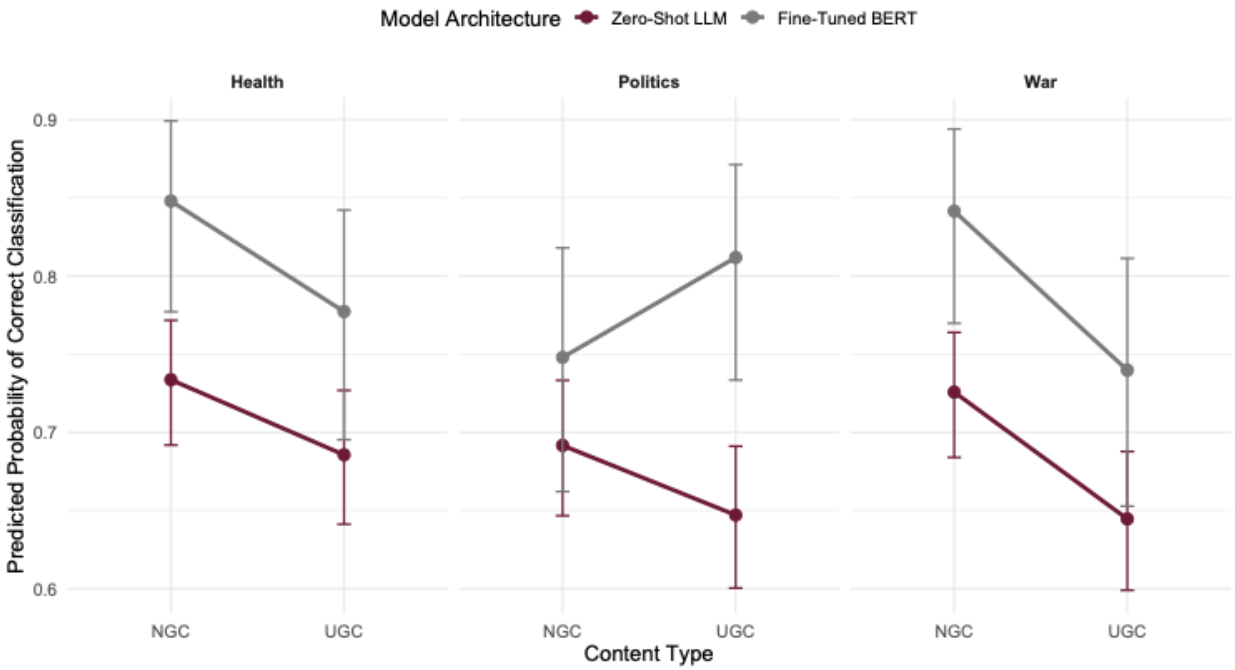
For the politics domain (relative to the Health reference category), the three-way interaction approached significance but did not meet the threshold (β = 0.811, SE = 0.4544, z = 1.785, p = .074). Exponentiation ($e^{0.811} \approx 2.25$) indicates that the architecture-specific difference in the UGC–NGC performance gap was approximately 125% greater in Politics relative to Health. This suggests that the extent to which the content type disparity differed between zero-shot LLMs and BERT-based models was amplified within the Politics domain compared to Health. However, this effect ultimately did not reach statistical significance.

The corresponding three-way interaction for the War domain was not statistically significant (β = -0.0117, SE = 0.4558, z = -0.026, p = .980), indicating no notable deviation in the architecture-specific UGC–NGC performance gap relative to the Health reference domain.

In Health and War, both architectures exhibited a comparable UGC-related performance penalty. However, in Politics, zero-shot LLMs showed a decline from NGC (0.692) to UGC (0.647), whereas fine-tuned BERT models reversed this pattern, with higher predicted accuracy for UGC (0.812) than for NGC (0.748).

**Figure 9**

*Predicted Classification Accuracy by Content Type and Architecture Across Domains*




*Note. 8* CI = 0.95

Overall, these findings suggest that although the architecture-specific content type disparity varies somewhat across domains, there is no statistically significant three-way moderation effect at the selected significance thresholds ($\alpha = 0.05$). The results, therefore, provide limited evidence that domain meaningfully alters how content-type vulnerabilities differ between zero-shot LLMs and fine-tuned BERT-based models.

## 4.3 Summary

This chapter presented the empirical results of the thesis using GLMM to examine misinformation detection performance across content type, domain, prompting condition, and model architecture. Section 4.1 demonstrated that content source type significantly influenced

LLM classification accuracy under reference conditions, with a UGC-related performance disadvantage that varied across domains and prompting strategies. A significant content type × domain interaction indicated that the UGC–NGC gap differed across Health, Politics, and War. Our significant three-way interaction showed that the prompting condition further moderated this disparity.

Section 4.2 extended the analysis to compare zero-shot LLMs and fine-tuned BERT-based models. While fine-tuned BERT-based models showed a significant overall accuracy advantage, there was limited evidence to suggest that the architecture meaningfully altered content type or domain-based performance disparities. Taken together, the results indicate that misinformation detection performance is shaped by structured interactions among the type of communication model reflected in social media content, topical domain, prompting strategy, and model architecture rather than by simple main effects alone.

# 5. Discussion

The findings presented in Chapter 4 show consistent differences in misinformation detection performance across content types, topical domains, prompting strategies, and model architectures. To interpret these results, this thesis draws on James Carey's (2009) transmission–ritual dichotomy, which contrasts communication as the efficient transmission of messages with communication as a symbolic process of community construction and meaning-making. We use Carey's framework to contextualize the observed performance differences between UGC and NGC, and to explain why prompting conditions, domains, and model architectures may moderate classification accuracy. The chapter concludes by outlining this study's contributions to FND research and discussing its implications for dataset design and evaluation practices.

## 5.1 The UGC-NGC performance disparity

This section directly addresses the first research question (RQ1) of this thesis, which asked whether misinformation detection systems exhibit differential classification performance across UGC and NGC. The GLMM results provide conditional evidence supporting the hypothesis that content source type significantly influences classification accuracy in misinformation detection systems.

In the LLM model (Section 4.1), a significant main effect of content type was observed under the reference condition (Health domain, few-shot prompting), with lower odds of correct classification for UGC relative to NGC. Model-adjusted predicted probabilities further indicated a consistent pattern in which UGC accuracy was lower than NGC accuracy across most modeled conditions. However, given the presence of higher-order interactions, this content-type effect

should be interpreted as conditional rather than universal. Nevertheless, the direction of the disparity was stable across model-adjusted predictions.

In the architecture comparison model (Section 4.2), both zero-shot LLMs and fine-tuned BERT-based models exhibited lower accuracy on UGC than NGC within the reference domain. Although the architecture × content type interaction was not statistically significant, predicted probabilities showed parallel declines across model families, indicating that the UGC-related performance penalty was broadly shared.

### *5.1.1 Transmission-oriented alignment and the NGC advantage*

Carey's (2009) transmission-ritual dichotomy provides a useful lens for interpreting the performance disparity outlined above. Transmission-oriented communication prioritizes clarity, accuracy, and efficient message transfer (Carey, 2009; Shannon & Weaver, 1949). Misinformation detection systems and LLMs in general are commonly evaluated using similar criteria, privileging rapid and consistent delivery of accurate outputs.

Institutional journalism similarly operates within transmission-oriented norms; it emphasizes verification, epistemic authority, structural consistency, and timeliness. A successful news organization breaks stories quickly while providing the public with accurate and up-to-date information. News headlines distributed via social media, print journalism, or broadcast media are crafted to convey information concisely and clearly. As a result, journalistic communication tends to follow standardized reporting that is structurally and linguistically distinct from discourse within online communities.

Recent legal and public debates surrounding LLM training data—for example, the lawsuit filed by The New York Times against OpenAI and Microsoft (New York Times,

2023)—reveal the extent to which news publications are represented in model pre-training corpora. While this is not direct evidence of model bias, it suggests that models encode the linguistic markers of institutional journalism during both training and evaluation.

The transmission view of communication shares values that also underpin both journalistic practice and model evaluation frameworks; this thesis therefore argues that such alignment advantages NGC in misinformation detection tasks.

Our experimental findings support this interpretation. The significant effect of content type observed in the LLM GLMM suggests that models perform better when the input reflects the values of the transmission model of communication. Under the reference condition, the odds of correctly classifying UGC were approximately half those of NGC. Although this effect was moderated by domain and prompting condition, the overall direction of predicted probabilities consistently favoured NGC.

### *5.1.2 UGC error-rates and ritual-oriented communication*

A closer examination of LM error rates provides further evidence of the structural misalignment between model assumptions and ritual-oriented communication. As shown in Table 4, LLMs exhibited higher false-positive and false-negative rates when evaluating UGC relative to NGC. UGC was more frequently misclassified as misinformation (False positive rate = 0.320 vs. 0.266), and misinformation embedded in UGC was more likely to be misclassified as true (False negative rate = 0.270 vs. 0.238). These classification patterns indicate that LLMs struggle more to reliably interpret features characteristic of ritual-oriented communication altogether, resulting in elevated error rates in both directions.

**Table 4**

*LLM Error Rates by Content Type*

| Content Type | False Positive Rate | False Negative Rate |
|---|---|---|
| NGC | 0.266 | 0.238 |
| UGC | 0.32 | 0.27 |

*Note. 9* False positive rate = proportion of true claims misclassified as misinformation. False negative rate = proportion of misinformation claims misclassified as true.

On the other hand, fine-tuned BERT models displayed a more directional asymmetry in their error-rate patterns (see Table 5). Unlike LLMs, which perform classification through prompting, BERT-based models are directly optimized to separate labeled categories during training. This means they learn to draw boundaries between "true" and "misinformation" based on patterns in their labeled training data. In our experiments, BERT-based models were substantially more likely to flag UGC as false misinformation than NGC (False positive rate = 0.278 vs. 0.169), while they were more likely to misclassify misinformation embedded in NGC as true (False negative rate = 0.274 vs. 0.229). This pattern suggests that when models are trained to discriminate between classes, they may lean towards the familiarity of the linguistic and structural cues associated with institutional journalism and mistakenly encode them as signals of credibility. As a result, transmission-oriented discourse may be implicitly granted epistemic privilege, while the many features characteristic of ritual-oriented communication are more readily penalized.

**Table 5**

*BERT error rates by content type*

| Content Type | False Positive Rate | False Negative Rate |
|---|---|---|
| NGC | 0.169 | 0.274 |
| UGC | 0.278 | 0.229 |

*Note. 10* False positive rate = proportion of true claims misclassified as misinformation. False negative rate = proportion of misinformation claims misclassified as true.

## 5.2 Prompting, model architecture and conditional gap attenuation

This section examines how different training paradigms influence LM performance in detecting misinformation across UGC and NGC (RQ3). Here, we discuss how prompting conditions influence performance gap disparities across content types and reflect on the implications of these findings for future FND research.

### *5.2.1 When few-shot learning is effective*

As seen in Chapter 4, the LLM GLMM revealed a significant three-way interaction among content type, domain, and prompting condition. A closer inspection of the model suggests that the NGC-UGC performance can be lessened through few-shot learning, but only in certain contexts. In the Health domain, few-shot prompting amplified rather than reduced the predicted probability gap between NGC and UGC. Few-shot prompted LLMs showed a predicted probability gap between NGC and UGC of 10.8 percentage points (0.877 vs. 0.769), compared to 5.9 percentage points under zero-shot prompting (0.758 vs. 0.699). These statistics show that

although few-shot prompting is effective at improving overall classification accuracy relative to the zero-shot baseline, it fails to improve performance consistently across content in the Health domain.

However, this effect did not generalize to the Politics and War domains. In Politics, zero-shot prompting produced a 5.5 percentage-point gap (0.748 vs. 0.693), whereas few-shot prompting resulted in near convergence between NGC and UGC (0.823 vs. 0.849; gap = −2.6 percentage points). A similar, though smaller, reduction in disparity was observed in the War domain, where the zero-shot gap was 8.8 percentage points (0.740 vs. 0.6252), while the few-shot gap narrowed to 4.9 percentage points (0.810 vs. 0.761).

These findings demonstrate that few-shot prompting is an effective approach for misinformation detection in theory, especially if the focus is on increasing overall classification accuracy. In practice, however, the success of few-shot prompting is highly conditional. The magnitude and direction of the NGC–UGC gap varied substantially across domains. Even where overall accuracy increased, these gains were not uniformly distributed across content types. Therefore, few-shot learning does not consistently resolve structural disparities in performance. Instead, its effects depend on the interaction between communicative form and topical domain. When considering the application of misinformation detection systems to social media content, these patterns demonstrate the importance of evaluation techniques that improve overall accuracy and account for how performance gains are distributed across different communication styles. Otherwise, improving absolute accuracy risks reinforcing persistent performance disparities across communication modes.

### *5.2.2 Model architecture and baseline performance*

RQ4 assessed whether architecture moderates content type and domain effects. Fine-tuned BERT models demonstrated significantly higher baseline accuracy than zero-shot LLMs, with approximately twice the odds of correct classification in the reference condition.

However, interactions involving architecture were largely non-significant, suggesting that although BERT-based models can achieve higher overall performance, they do not fundamentally alter the structure of content-type disparities relative to LLMs. We also observed that overall improvements in predicted probability of correct classification do not eliminate the UGC-related performance reduction in BERT-based models. This distinction is especially important for FND research, where researchers often frame the highest accuracy scores as the ultimate goal.

### *5.2.3 Implications for model evaluation and dataset design*

These findings suggest that optimizing solely for overall accuracy may inadvertently lead models to become better at detecting transmission-oriented communication patterns and mass media rituals, rather than recognizing instances of misinformation. The results of our few-shot learning experiment demonstrate that performance gaps are not inevitable and that more consistent detection across types of communication is achievable with training samples that represent the range of communication styles found on social media. Thus, addressing UGC-NGC disparities requires intentional dataset design that encompasses both transmission- and ritual-oriented communication patterns. Without such measures, models may become more accurate overall while continuing to underperform on community-oriented content, resulting in uneven detection performance across communication styles.

## 5.3 Domain-specific communicative dynamics

This section interprets our findings of domain-specific patterns through Carey's (2009) transmission-ritual framework, discussing how communicative dynamics differ across Health, Politics, and War. We argue that variation in model performance reflects differences in how each domain structures authority, community meaning-making, and stylistic convergence between institutional and participatory discourse.

### *5.3.1 Communication rituals and health information behaviours*

In the Health domain, the UGC disadvantage was evident across conditions (see Table 6). This pattern may partly reflect the credibility cues users prioritize when seeking health-related information, which closely align with transmission-oriented norms. As a result, health journalism emphasizes signals of institutional authority and evidence-based reporting—such as source linking and factual clarity—over opinion-driven or experiential content.

At the same time, user-generated health communication often follows a different trajectory, particularly during prolonged public health crises that require sustained sense-making (Dervin, 1983). Prior work shows that as users accumulate information, their participation increasingly incorporates personal opinions and emotional language (Heverin & Zach, 2012; Vos & Buckner, 2015). In the context of the multi-year COVID-19 pandemic, UGC frequently reflected attempts to process uncertainty, share lived experiences, and negotiate meaning collectively, rather than to transmit authoritative medical information. The result is a user-based health discourse that diverges substantially from the conventions of institutional health reporting.

**Table 6**

*Predicted Probability for Correct Classification in the Health Domain*

| Content type | Prompting condition | Predicted probability of correct classification |
|---|---|---|
| NGC | Few-shot | 0.877 |
| UGC | Few-shot | 0.769 |
| NGC | Zero-shot | 0.758 |
| UGC | Zero-shot | 0.699 |
| NGC | BERT | 0.785 |
| UGC | BERT | 0.728 |

*Note.* 11 CI = 0.95

Carey's (2009) ritual view of communication helps explain this divergence. In extended health crises, the purpose of communication is many things. It serves to convey information but also to sustain community, share anxieties, and reinforce social belonging during periods of isolation and uncertainty. These ritual functions are especially pronounced in UGC, where expressive and affective communication plays a central role. News organizations, by contrast, continue to serve as trusted information brokers, maintaining transmission-oriented norms even as public discourse becomes increasingly experiential.

The divergence between ritual-oriented UGC and transmission-oriented NGC offers a potential explanation for the observed performance gap across models in the Health domain. LMs' pre-training corpora likely bias them toward institutional credibility cues, making them better equipped to process content that reflect the values underpinning the transmission model of communication than discourse grounded in sense-making and emotional expression. Therefore, it

makes sense that even with the addition of selectively curated labeled examples, LMs perform better on news-generated Health content than on user-generated Health communication.

### *5.3.2 Communication rituals and political information behaviours*

In Politics, predicted probabilities suggested a more complex pattern. In the architecture model, fine-tuned BERT-based models showed higher predicted accuracy for UGC (0.812) than for NGC (0.748), and a similar reversal was observed for LLMs in the few-shot condition, where they outperformed on political UGC (0.849) compared to NGC (0.805). One possible explanation for this slope inversion is that contemporary political journalism increasingly targets ideologically segmented audiences rather than a politically neutral public. Thus, political news content often moves beyond mere information transmission to incorporate cues that resonate with identity-driven forms of engagement.

From a user experience perspective, political content is often embedded in environments characterized by strong partisan alignment and community affiliation. Major news organizations have become closely associated with specific ideological audiences. For example, Fox News is closely associated with Republican-aligned viewers, while NBC and MSNBC orient toward Democratic audiences. This creates traditional media ecosystems that parallel the ideological echo chambers facilitated by curated content delivery on social media. Within these environments, both institutional and user-generated political content serve as sites of identity signalling, boundary maintenance, and group affirmation, thereby reducing the prominent distinction between UGC and NGC.

Carey's (2009) ritual view of communication helps explain why this alignment in audience orientation produces convergence across communicative forms. When political discourse

prioritizes the performance of ideological identity and the reinforcement of community belonging, communication operates less as neutral information transmission and more as a ritualized social practice. Under these conditions, institutional political journalism adopts participatory and performative features traditionally associated with UGC, while user-generated political discourse mirrors the stylistic and thematic conventions of partisan news media.

This takeaway offers a plausible explanation for why few-shot prompted LLMs and fine-tuned BERT models perform more effectively on political UGC than on NGC. As models become attuned to ritualized signals of affiliation that permeate political communication across source types, the expected distinction between institutional and user-generated content becomes less salient, allowing UGC to achieve equal or greater model performance in the Politics domain.

### *5.3.3 Communication rituals and war information behaviours*

In the War domain, content-type differences were comparatively small, and interactions were not statistically significant. This narrowing performance gap may stem from the inherently opaque nature of information transfer in conflict-related reporting. War journalism operates under conditions of restricted access, official censorship, and propaganda—all factors that complicate verification and obscure clear distinctions between authoritative transmission and participatory ritual communication. When institutional sources face such constraints, the epistemic authority that typically differentiates NGC from UGC becomes attenuated.

An additional factor that may contribute to convergence in UGC-NGC predicted classification accuracy is the relative experiential distance from war and conflict for many users of the sampled platforms. Truth Social is officially supported only in North American countries (Forbes, 2025). While Bluesky is available globally, our sample was limited to English-language

content, which means it predominantly reflects Western users' experiences. Unlike domestic political or health-related content, which often directly affects users' daily lives, identities, and material conditions, war and conflict are frequently experienced as geographically and socially remote phenomena for a substantial portion of the sampled users. This does not imply a lack of concern or engagement, nor does it discount the presence of users with direct experience of conflict. Rather, it suggests that user-generated war-related content may be less rooted in personal stakes or lived experience than content in domains with more immediate consequences.

We can use Carey's (2009) ritual view of communication to explain why this geographic distance could account for the balanced performance between UGC and NGC in the War domain. Because many Western users do not directly identify with or experience the active conflicts they discuss, the ritual aspects of communication—such as signalling identity and establishing community—may be less pronounced in UGC. In these cases, users may instead adopt communicative strategies that resemble institutional reporting, emphasizing factual updates or secondary information to fill the gaps left by lower ritual engagement. Similarly, war journalism may incorporate content that reflects the ritual view of communication, such as eyewitness accounts and statements of solidarity, to fill gaps created by restricted access or limited information. As a result, both UGC and NGC combine transmission and ritual functions, which reduces stylistic divergence between the two content types.

This convergence in communicative function helps explain the narrower performance gap observed across models in the War domain and may extend to other contexts where information opacity diminishes the traditional epistemic advantages of institutional sources. Under such

conditions, few-shot models can detect misinformation patterns in content that simultaneously serves both transmission- and ritual-oriented roles.

## 5.4 Limitations

This section discusses the methodological and structural limitations of this study, highlighting factors that influenced the ability to address research questions, particularly RQ2, and the generalizability of findings.

### *5.4.1 Methodological limitations*

A key methodological limitation concerns the selection of research samples and data. This thesis explicitly aimed to address a gap in existing research by examining LM misinformation detection performance across UGC and NGC. A simulation-based power analysis using the *simr* package in R indicated that a sample of n = 1,032 claims (516 per content type across nine LMs) provided 95% power to detect content source effects.

However, when both content-source and topical-domain stratified analyses were conducted, sample sizes within each condition were substantially reduced, limiting power to detect smaller domain-specific effects or interactions. Power analyses for these stratified conditions fell below the conventional 80% threshold. Therefore, non-significant findings regarding domain effects or interactions should be interpreted with caution because insufficient sample sizes may have obscured true effects.

In addition to these limitations, computational and funding constraints limited the study to freely accessible and relatively small models, preventing local evaluation of large frontier architectures. This constrains the generalizability of our findings to highly parameterized or resource-intensive models.

### *5.4.2 Researcher and structural limitations*

Another methodological limitation is the lack of prior research that explicitly distinguishes between UGC and NGC in misinformation-detection tasks. While previous work has examined misinformation across social media and news separately, none (to the best of my knowledge) treat content source type as a primary analytical variable. Thus, this thesis operates in a relatively underexplored research space, which limits opportunities for direct comparison with established benchmarks and methodological guidance, which are particularly valuable in master-level research.

Time constraints inherent to the Master's thesis timeline shaped the study's design. Data were drawn from unlabeled, publicly available English-language social media datasets. This limited the demographic, linguistic, cultural, and temporal diversity of our sample. While these choices ensured analytical rigour, they limit the generalizability of the findings beyond the sampled platforms and content types.

Finally, although interpretations were grounded in Carey's (2009) framework, any analysis of communication practices risks being shaped by cultural assumptions. The dataset was limited to English-language content from Bluesky Social and Truth Social, which likely skewed the sample toward Western users and perspectives. This limitation is especially relevant for domains like War, where lived experience, cultural context, and geographic proximity may meaningfully shape communication practices. As a result, our findings may not generalize to non-English or non-Western social media contexts.

# 6. Conclusion

This thesis set out to examine whether misinformation detection systems perform differently on UGC and NGC and to understand why such disparities emerge when they do. Using a GLMM to account for claim-level variability and interaction effects, the findings demonstrate that content source is a significant (but conditional) determinant of classification performance. Rather than revealing a uniform bias, the results show that the magnitude and direction of UGC-NGC disparities depend on the prompting condition, topical domain, and model architecture.

Drawing on James Carey's (2009) transmission-ritual framework, we argue that content-specific performance disparities reflect the degree of alignment between communicative assumptions embedded in LMs and the linguistic patterns of content produced by groups that align with a given communicative tradition. That is, communicative function operates on LM performance indirectly, through the values and norms of a given communicative model that manifest as distinct linguistic patterns that models are trained to recognize and privilege over others.

NGC and the transmission model of communication share a fundamental conception of communication as an act of information delivery, and it is this shared understanding that produces convergent values of accuracy, timeliness and clarity. These values are also inherent to the training process of LMs, whose optimization objectives reward factual precision and informational coherence. The result is a tripartite alignment between the transmission model of communication, NGC production norms, and LM training assumptions.

In contrast, the ritual model of communication conceives communication as a shared experience that constructs and maintains social reality. This understanding is reflected in UGC, where users prioritize identity performance and community maintenance, producing linguistic patterns shaped by dialogic convention, affective expression, and in-group references. These features sit outside the values privileged by NGC and LM training objectives. The result is a misalignment between ritual-oriented communicative values and LM training assumptions, resulting in degraded model performance and elevated misclassification rates.

The findings from this study suggest that these disparities are neither insurmountable nor uniform. The few-shot prompting condition significantly reduced the UGC-NGC performance gap observed in the zero-shot condition in the Health and War domains, but not in the Politics. This finding suggests that limited exposure to labeled examples can recalibrate model decision boundaries, improving adaptability to diverse communicative forms. However, the presence of significant interaction effects and higher-order interactions means that the effectiveness of few-shot prompting is highly context-dependent.

At the same time, the fine-tuned baseline models illustrate an important trade-off. We found that supervised training substantially increased overall classification accuracy relative to zero-shot and few-shot LLMs. However, this improvement did not resolve disparities in classification performance across content types. In some conditions, we observed that fine-tuning preserved or even amplified NGC's advantages. This divergence between absolute accuracy and performance consistency highlights a central tension in FND: optimizing for overall accuracy alone may inadvertently reduce model sensitivity to participatory or community-oriented communicative spaces that are often the focus of misinformation interventions.

Taken together, these findings demonstrate that aggregate performance metrics can mask meaningful interaction effects. Models may appear effective at the macro level while exhibiting conditional weaknesses in precisely the environments where automated moderation tools are hoping to be deployed. By centring content source type as an analytical variable within a mixed-effects framework, this thesis exposes a structural blind spot in prevailing current FND benchmarking and evaluation practices.

## 6.1 Implications for FND

This section outlines the practical implications of this study's findings for the design, training, and evaluation of future FND systems. Our findings suggest that meaningful progress toward deployable systems requires attention to content source type, training paradigm selection, and evaluation practices that are sensitive to the communicative diversity of real-world online discourse.

Datasets that combine UGC and NGC without explicitly modeling content source type can generate misleading performance estimates. In such cases, LMs may disproportionately rely on institutional credibility cues that do not generalize to community discourse. For example, a model trained on mixed-source data may achieve high overall accuracy by correctly classifying professionally produced news content, while underperforming on user-generated posts that lack clear authority signals. Dataset construction should therefore stratify by content source type and report disaggregated performance estimates to ensure that benchmark results reflect the full communicative range of social media environments.

Furthermore, the domain-dependent patterns observed in the few-shot condition highlight the necessity of domain-sensitive evaluation. Evaluation protocols that collapse across domains

risk obscuring meaningful performance variations. Different topics attract distinct audiences that produce varied discourse norms and authority structures. These variations influence how misinformation is produced and how FND systems respond to it. Therefore, domain-stratified reporting should be adopted as standard practice to enable researchers to identify topics where systems are reliable and where they need improvement.

Finally, our findings demonstrate that different training paradigms lead to distinct LM performance patterns. Few-shot learning improved adaptability to ritual-oriented content, whereas fine-tuning improved overall performance but increased disparities across content source types. These results suggest that researchers should carefully consider the communicative characteristics of the target deployment environment when selecting a training approach. For instance, in deployment settings where content is heterogenous, few-shot approaches may be preferable for maintaining contextual sensitivity, whereas fine-tuning may be better suited to more standardized environments, where aggregate metrics are less likely to obscure meaningful variation.

## 6.2 Future directions

Future research should build on this work by expanding both the scale and scope of analysis. A larger dataset would enable greater statistical power for detecting higher-order effects and more granular analyses of how communicative form interacts with model architecture.

Extending this work to multilingual, multimodal, and cross-cultural corpora would allow researchers to examine how ritual communication varies across contexts and how such variation influences automated classification.

Future studies should also explore how different training paradigms affect both overall accuracy and consistency of model performance across content types. Hybrid or adaptive training strategies may provide a path toward models that generalize effectively while maintaining sensitivity to diverse communication patterns. Emerging frameworks such as DSPy, which enable the optimization of prompting pipelines and task-specific reasoning strategies, offer a promising direction for improving model performance across heterogeneous content types (DSPy, 2025). By moving beyond static prompt design and toward dynamically optimized workflows, such approaches may help mitigate performance disparities observed between UGC and NGC. Finally, evaluating a broader range of models, including large-scale frontier architectures and systems trained under varied pre-training regimes, could clarify whether the patterns observed here persist at scale or whether increased model capacity mitigates existing biases.

## 6.3 Concluding remarks

This thesis argues that automated misinformation detection depends in part on communicative alignment. Using a GLMM to analyze performance, the findings reveal disparities arising from interactions among content source type, domain and training paradigm.

Therefore, addressing misinformation in social media environments requires evaluation strategies that are sensitive to linguistic patterns reflecting both the transmission and ritual models of communication. Toward this end, achieving deployable systems depends jointly on advances in algorithm design alongside careful dataset construction and rigorous evaluation practices.

By bringing Carey's transmission–ritual framework into dialogue with GLMM-based modeling, this thesis presents a more socially attuned approach to FND—one that recognizes how differing perceptions of the purpose of online communication shape linguistic patterns and, ultimately, system performance.

# Bibliography


Ahmed, S. (2021). Combining text classification and fact checking to detect fake news. *Electronic Theses and Dissertations.* https://pubblicazioni.unicam.it/retrieve/ca2cbed8-819f-4b08-a8c2 01731cf2e9ad/26_08_22%20Ahmed%20Sajjad-Thesis-Final_revised%20%281%29.pdf

Ajao, O., Bhowmik, D., & Zargari, S. (2019). Sentiment-aware fake news detection on online social networks. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2507–2511. https://doi.org/10.1109/ICASSP.2019.8683170

Al-Ash, H. S., & Wibowo, W. C. (2018). Fake news identification characteristics using named entity recognition and phrase detection. *Proceedings of the 2018 International Conference on Advanced Computer Science and Information Systems*, 12–17. https://doi.org./10.1109/ICITEED.2018.8534898

Alghamdi, J., Lin, Y., & Luo, S. (2024). Cross-domain fake news detection using a prompt-based approach. *Future Internet, 16*(8). https://doi.org/10.3390/fi16080286

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives, 31*(2), 211-36. https://doi.org/10.1257/jep.31.2.211

Alhabash, S., & Ma, M. (2017). A tale of four platforms: Motivations and uses of Facebook, Instagram, and Snapchat among college students? *Social Media + Society, 3*(1). https://doi.org/10.1177/2056305117691544

Alnabhan, M.Q. (2025). *Advancing cross-domain fake news detection: Enhanced models to improve generalization and tackle the class imbalance problem*. [Doctoral dissertation, University of Ottawa]. https://hdl.handle.net/10393/50256

Alnabhan, M. Q., & Branco, P. (2024). BERTGuard: Two-tiered multi-domain fake news detection with class imbalance mitigation. *Big Data and Cognitive Computing, 8*(8). https://doi.org/10.3390/bdcc8080093

Anderson, M., Faverio, M., & Gottfried, J. (2023, December 11). Teens, social media and technology 2023. *Pew Research Center.* https://www.pewresearch.org/internet/2023/12/11/teens-social-media-and-technology-2023/

Alquadi, B.S., Alushibany, S.A., Yousafzai, S.M., Alzu'bi, S., Alsekait, D.M., & AbdElminaam, D.S. (2025). Transfer learning driven fake news detection and classification using large language models. *Scientific Reports, 15*(28490). https://doi.org/10.1038/s41598-025-10670-2

Authoritas. (2025, January 27). *AI Overview: user intent research*. Authoritas. https://www.authoritas.com/blog/ai-overview-user-intent-research

Avram, M., Micallef, N., Patil, S., & Menczer, F. (2020). Exposure to social engagement metrics increases vulnerability to misinformation. *Misinformation Review, 1*(5). https://doi.org/10.37016/mr-2020-033

Barney, D., Coleman, G., Ross, C., Sterne, J., & Tembeck, T. (2016). *The participatory condition in the digital age*. University of Minnesota Press. https://muse.jhu.edu/book/48363

Bastos, M., & Tuters, M. (2023). Meaningful disinformation: Narrative rituals and affective folktales. *Big Data & Society, 10*(2). https://doi.org/10.1177/20539517231215361

Baum, L. E., & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics, 37*(6), 1554–1563. https://doi.org/10.1214/aoms/1177699147

Benamira, A., Devillers, B., Lesot, E., Ray, A. K., Saadi, M., & Malliaros, F. D. (2019). Semi-supervised learning and graph neural networks for fake news detection. *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 568–569. https://doi.org/10.1145/3341161.3342958

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bates, M. (1989). The design of browsing and berrypicking: Techniques for the online search interface. *Online Review, 13*(5), 407–424. https://pages.gseis.ucla.edu/faculty/bates/articles/berrypicking.pdf

Benny, J.J. (2024). *Knowledge-informed fake news detection using large language models.* [Masters dissertation]. Electronic Theses and Dissertations. 9195. https://scholar.uwindsor.ca/etd/9195

Bogert, E., Schecter, A., & Watson, R. T. (2021). Humans rely more on algorithms than social influence as a task becomes more difficult. *Scientific Reports, 11*(1), Article 8028. https://doi.org/10.1038/s41598-021-87480-9

Borges do Nascimento, I. J., Pizarro, A. B., Almeida, J. M., Azzopardi-Muscat, N., Gonçalves, M. A., Björklund, M., & Novillo-Ortiz, D. (2022*). Infodemics and health misinformation:*

*a systematic review of reviews. Bulletin of the World Health Organization, 100*(9), 544–561. https://doi.org/10.2471/BLT.21.287654

Bradshaw, S., & Howard, P.N. (2018). The global organization of social media disinformation campaigns. *Journal of International Affairs, 71*(1.5), 23–32. https://www.jstor.org/stable/26508115

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems,* 33, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

Bolker, B., Robinson, D., & others. (2024). *broom.mixed: Tidying methods for mixed models (R package)*. https://CRAN.R-project.org/package=broom.mixed

Boot, A.B., Dijkstra, K. & Zwaan, R.A. (2021). The processing and evaluation of news content on social media is influenced by peer-user commentary. *Humanities and Social Sciences Communications 8*(209). https://doi.org/10.1057/s41599-021-00889-5

Cacioppo, J.T., Hawkley, L.C., Norman, G.J., & Bernston, G.G. (2011). Social isolation. *Annals of the New York Academy of Sciences, 1231*(1), 17–22. https://doi.org/10.1111/j.1749-6632.2011.06028.x

Cao, C., Sang, J., Arora, R., Chen, D., Kloosterman, R., Cecere, M., Gorla, J., Saleh, R., Drennan, I., Teja, B., Fehlings, M., Ronksley, P., Leung, A. A., Weisz, D. E., Ware, H., Whelan, M., Emerson, D. B., Arora, R. K., & Bobrovitz, N. (2025). Development of prompt templates for large language model-driven screening in systematic reviews.

*Annals of internal medicine, 178*(3), 389–401. https://doi.org/10.7326/ANNALS-24-02189

Caramancion, K. M. (2023). News verifiers showdown: A comparative performance evaluation of ChatGPT 3.5, ChatGPT 4.0, Bing AI, and Bard in news fact-checking. *In 2023 IEEE Future Networks World Forum (FNWF)*, 1-6. IEEE. https://doi.org/10.48550/arXiv.2306.17176

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on Twitter. *Proceedings of the 20th international conference on World wide web*, 675–684. https://doi.org/10.1145/1963405.1963500

Carey, J. W. (2009). *Communication as culture: Essays on media and society* (Rev. ed.). Routledge. ISBN: 9780415989756

Carey, J. W., & Adam, G. S. (2008). *Communication as culture: Essays on media and society* (Rev. 2nd ed.). Routledge. ISBN 9780415989763

Chapekis, A., & Lieb, A. (2025, July 22). *Google users are less likely to click on links when an AI summary appears in the results.* Pew Research Center. https://www.pewresearch.org/short-reads/2025/07/22/google-users-are-less-likely-to-click-on-links-when-an-ai-summary-appears-in-the-results/

Chern, E., Chern, S., Chen, S., Yuan, W., Feng, K., Zhou, C., He, J., Neubig, G., & Liu, P. (2023). FacTool: Factuality detection in generative ai - a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint*. arXiv, abs/2307.13528.

Cheung, T. H., & Lam, K.M. (2023). FactLLaMA: Optimizing instruction-following language models with external knowledge for automated fact-checking. *arXiv preprint.* https://doi.org/10.48550/arXiv.2309.00240

Choi, A. S., Akter, S. S., Singh, J. P., & Anastasopoulos, A. (2024). The LLM effect: Are humans truly using LLMs, or are they being influenced by them instead? *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 22032–22054. https://doi.org/10.18653/v1/2024.emnlp-main.1230

Choudhry, A., Khatri, I., Chakraborty, A., Vishwakarma, D.K., & Prasad, M. (2022). Emotion-guided cross-domain fake news detection using adversarial domain adaptation. arXiv preprint, doi:10.48550/arXiv.2211.13718

Cinelli, M., Morales, G.D.F., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences. 118*(9). https://doi.org/10.1073/pnas.2023301118

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science, 385*(6714). https://doi.org/10.1126/science.adq1814

Cui, A., Zhang, M., Liu, Y., Ma, S. (2011). Emotion Tokens: Bridging the gap among multilingual twitter sentiment analysis. *In: Salem, M.V.M., Shaalan, K., Oroumchian, F., Shakery, A., Khelalfa, H. (eds) Information Retrieval Technology. AIRS 2011. Lecture Notes in Computer Science, 7097*. https://doi.org/10.1007/978-3-642-25631-8_22

Dadkhah, S., Zhang, X., Weismann, A. G., Firouzi, A., & Ghorbani, A.A. (2023). The largest social media ground-truth dataset for real/fake content: TruthSeeker. *In IEEE*

*Transactions on Computational Social Systems, 99*. 1-15. https://www.unb.ca/cic/datasets/truthseeker-2023.html

Darcy, O. (2019, March 22). *How Twitter's algorithm is amplifying extreme political rhetoric.* CNN. https://www.cnn.com/2019/03/22/tech/twitter-algorithm-political-rhetoric/index.html

Dervin, B. (1998). Sense-making theory and practice: An overview of user interests in knowledge seeking and use. *Journal of Knowledge Management, 2*(2), 36-46. https://doi.org/10.1108/13673279810249369

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *North American Chapter of the Association for Computational Linguistics.* https://doi.org/doi:10.18653/v1/N19-1423

Dey, A., Rafi, R.Z., Parash, S.H., Arko, S.K., Chakrabarty, A. (2018). Fake news pattern recognition using linguistic analysis. *Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, 305–309. IEEE. https://doi.org/10.1109/ICIEV.2018.8641018

Diakopoulos, N. (2014). Algorithmic accountability reporting: On the investigation of black boxes (Tow Center for Digital Journalism). *Columbia University.* https://doi.org/10.7916/D8ZK5TW2

DSPy. (2025). *DSPy Optimizers (formerly Teleprompters).* https://dspy.ai/learn/optimization/optimizers/

Eisenstein, J. (2019). *Introduction to natural language processing.* MIT Press.

Element Labs. (2026). *LM Studio.* [Computer software]. https://lmstudio.ai

Ellis, D. (1989). A behavioural approach to information retrieval system design. *Journal of Documentation, 45*(3), 171–212. https://doi.org/10.1108/eb026843

Emtage, A. (1990). *Archie.* McGill University.

Failla, A., & Rossetti, G. (2024). "I'm in the Bluesky Tonight": Insights from a year worth of social data. *PLoS ONE, 19*(11). https://doi.org/10.1371/journal.pone.0310330

Fernandez, M., Bellogín, A., & Cantador, I. (2024). Analysing the effect of recommendation algorithms on the spread of misinformation. *Proceedings of the 16th ACM Web Science Conference (WEBSCI '24)*. 159–169. https://doi.org/10.1145/3614419.3644003

Fernández-Pichel, M., Pichel, J.C. & Losada, D.E. (2025). Evaluating search engines and large language models for answering health questions. *Digital Medicine* 8(153) https://doi.org/10.1038/s41746-025-01546-w

Gerard, P., Botzer, N., & Weninger, T. (2023). Truth social dataset [Dataset and description]. *In Proceedings of the International AAAI Conference on Web and Social Media (ICWSM).* https://arxiv.org/abs/2303.11240

González-Silot, S., Montoro-Montarroso, A., Cámara, E. M., & Gómez-Romero, J. (2025). Enhancing disinformation detection with explainable AI and named entity replacement. *arXiv preprint*. arXiv:2502.04863.

Google. (2024). *AI Overviews in Search*. https://search.google/ways-to-search/ai-overviews/

Google Cloud. (2026, April 1). *Prompt engineering: Overview and guide.* https://cloud.google.com/discover/what-is-prompt-engineering

Gunjal, S. (2021). *Tutorial: K-fold cross validation*. Kaggle. https://www.kaggle.com/satishgunjal/tutorial-k-fold-cross-validation

Guo, Z., Schlichtkrull, M., & Vlachos, A. (2021). A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics, 10*, 178-206.

He, P., Liu, X., Gao, J., & Chen, W. (2020). DeBERTa: Decoding-enhanced bert with disentangled attention. *arXiv preprint*. arXiv:2006.03654.

Heverin, T., & Zach, L. (2011). Use of microblogging for collective sense-making during violent crises: A study of three campus shootings. *Journal of the American Society for Information Science and Technology, 63*(1)*, 34-47.* https://doi.org/10.1002/asi.21685

Hoffman, T., Schölkopf, B., & Smola, A.J. (2008). Kernel methods in machine learning. *The Annals of Statistics, 36*(3), 1171-1220. https://doi.org/10.1214/009053607000000677

Hu, B., Sheng, Q., Cao, J., Shi, Y., Li, Y., Wang., & Qi., P. (2023). Learn over past, evolve for future: Forecasting temporal trends for fake news detection. *Proceedings of the 61st*

*Annual Meeting of the Association for Computational Linguistics, 5*. https://doi.org/10.18653/v1/2023.acl-industry.13

Iandola, F., Shaw, A., Krishna, R., & Keutzer, K. (2020, November). SqueezeBERT: What can computer vision teach NLP about efficient neural networks? *In Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, 124-135. https://doi.org/10.48550/arXiv.2006.11316

IBM. (n.d.). *What is classification in machine learning?* IBM Think. https://www.ibm.com/think/topics/classification-machine-learning

Jang, J., Ye, S., & Seo, M. (2023) Can large language models truly understand prompts? a case study with negated prompts. *arXiv preprint,* 52-62. https://doi.org/10.48550/arXiv.2209.12711

Jiang, Y., Li, X., Zhu, G., Li, H., Deng, J., Han, K., Shen, C., Shi, Q., & Zhang, R. (2023). 6G non-terrestrial networks enabled low-altitude economy: Opportunities and challenges. *arXiv preprint* arXiv:2311.09047.

Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications, 80*(8), 11765–11788. https://doi.org/10.1007/s11042-020-10183-2

Kennedy, S. (2024). Algorithmic recommender systems: Mitigating information overload and preserving spontaneous freedom. *American Philosophical Quarterly*, *61*(4). 327–338. https://doi.org/10.5406/21521123.61.4.03

Kim, M. G., Kim, M., Kim, J. H., & Kim, K. (2022). Fine-Tuning BERT Models to Classify Misinformation on Garlic and COVID-19 on Twitter. *International Journal of Environmental Research and Public Health, 19*(9), 5126. https://doi.org/10.3390/ijerph19095126

Kishimoto, K. & Fukushima, N. (2011). Use of anonymous web-communities and websites by medical consumers in Japan to research drug information. *Yakugaku Zasshi*, 131.

Kochkina, E., Liakata, M., & Zubiaga, A. (2018) *PHEME dataset for Rumour Detection and Veracity Classification.* https://doi.org/10.6084/m9.figshare.6392078.v1

Kotonya, N., & Toni, F. (2020). Explainable automated fact-checking for public health claims. *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 7740-7754. https://doi.org/10.48550/arXiv.2010.09926

Kugler, L. (2024). How today's recommender systems use machine learning to cater to your every whim. *Communications of the ACM.* https://doi.org/10.1145/3673426

Kulthau, C.C. (1991) Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science, 42*(5), 361–371. https://doi.org/10.1002/(SICI)1097-4571(199106)42:5<361::AID-ASI6>3.0.CO;2-#

Kung, P.N., & Peng, N. (2023). Do models really learn to follow instructions? An empirical study of instruction tuning. *In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2,* 1317–1328. https://doi.org/10.18653/v1/2023.acl-short.113

Langguth, J., Schroeder, D. T., Filkuková, P., Brenner, S., Phillips, J., & Pogorelov, K. (2023). COCO: an annotated Twitter dataset of COVID-19 conspiracy theories. J*ournal of Computational Social Science*, 1–42. https://doi.org/10.1007/s42001-023-00200-3

Lee, E. J., Kim, H. S., & Joo, M. H. (2023). Social media vs. mass media: Mitigating the suspicion of ulterior motives in public health communication. *Health Communication, 38*(11), 2450–2460. https://doi.org/10.1080/10410236.2022.2074781

Lerman, K., & Ghosh, R. (2010). Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. *Proceedings of 4th International AAAI Conference on Weblogs and Social Media.* https://doi.org/10.48550/arXiv.1003.2664

Lewandowsky, S., Gignac, G.E., & Oberauer, K. (2015) Correction: The role of conspiracist ideation and worldviews in predicting rejection of science. *PLOS ONE 10*(8): e0134773. https://doi.org/10.1371/journal.pone.0134773

Lewandowsky, S., Ecker, U.K., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the post-truth era. *Journal of Applied Research in Memory and Cognition*, *6*(4), 353-369. https://doi.org/10.1016/j.jarmac.2017.07.008

Lewin, K. (1947). Frontiers in group dynamics: ii. Channels of group life; social planning and action research. *Human Relations, 1*(2), 143-153. https://doi.org/10.1177/001872674700100201

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in neural information processing systems, 33*, 9459-9474. https://doi.org/10.48550/arXiv.2005.11401

Li, J., Li, G., Shi, Y., & Yu, Y. (2021). Cross-domain adaptive clustering for semi-supervised domain adaptation. a*rXiv preprint.* Doi 10.1109/CVPR46437.2021.00253

Li, Q., & Zhou, W. (2020). Connecting the dots between fact verification and fake news detection. *Proceedings of the 28th International Conference on Computational Linguistics,1820-1825.* https://doi.org/10.18653/v1/2020.coling-main.165

Liu, Y., & Wu, Y.-F. B. (2018). Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. *In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligen*ce, 354–361. https://doi.org/10.1609/aaai.v32i1.11268

Liu, H., Wang, W., Li, H., & Li, H. (2024). TELLER: A Trustworthy Framework for Explainable, Generalizable and Controllable Fake News Detection. *arXiv preprint*. arXiv:2402.07776

Liu, X., Nourbakhsh, A., Li, Q., Fang, R., & Shah, S. (2015) Real-time rumor debunking on twitter. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 1867–1870. https://doi.org/10.1145/2806416.2806651

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint*. arXiv:1907.11692.

Liu, Y., Zhu, J., Zhang, K., Tang, H., Zhang, Y., Liu, X., Liu, Q., & Chen, E. (2024). Detect, investigate, judge, and determine: A novel LLM-based framework for few-shot fake news detection. arXiv preprint. arXiv:2407.08952

Liu, Z., Yang, K., Xie, Q., Kock, C. de, Ananiadou, S., & Hovy, E. (2024). RAEmoLLM: Retrieval augmented LLMs for cross-domain misinformation detection using in-context learning based on emotional information. a*rXiv preprint.* https://doi.org/10.48550/arXiv.2406.11093

Liu, Z., Zhang, T., Yang, K., Thompson, P., Yu, Z., & Ananiadou, S. (2024). Emotion detection for misinformation: *A review. Information Fusion*, *107.* https://doi.org/10.1016/j.inffus.2024.102300

Lo, S.L. (2023). The CLEAR Path: A Framework for Enhancing Information Literacy through Prompt Engineering. *Journal of Academic Librarianship, 4.* https://doi.org/10.1016/j.acalib.2023.102720

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to *human judgment. Organizational* Behavior *and Human Decision Processes*, 151, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Lüdecke, D., Makowski, D., Waggoner, P., & Patil, I. (2026). *performance: Assessment of regression models performance* (R package). https://CRAN.R-project.org/package=performance

Ma, J., Gao, W., Wei, Z., Lu, Y., & Wong, K.-F. (2015). Detect rumors using time series of social context information on microblogging websites. *In Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, 1751–1754. https://doi.org/10.1145/2806416.2806607

Ma, J., Gao, W., & Wong, K. F. (2017, July). Detect rumors in microblog posts using propagation structure via kernel learning. *In Proceedings of the 55th Annual Meeting of*

*the Association for Computational Linguistics, 1*, 708-717. https://doi.org/10.18653/v1/P17-1066

Mannuru, N. R. , Mannuru, A. and Lund, B. (2024). Large Language Models (LLMs) as a tool to facilitate information seeking behavior. *Information Science Trends*, 1(3), 34-42. doi: 10.61186/ist.202401.01.15

McCulloch, W. S., & Pitts, W. (1944). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics, 5*(4), 115–133. https://doi.org/10.1007/BF02478259

Meta. (2025). *Transparency Center - Misinformation: Policy Details.* https://transparency.meta.com/policies/community-standards/misinformation

Milli, S., Carroll, M., Wang., Y., Pandey, S., Zhao, S., & Dragan, A.D. (2025). Engagement, user satisfaction, and the amplification of divisive content on social media. *PNAS Nexus, 4*(3). https://doi.org/10.1093/pnasnexus/pgaf062

Mirbabaie, M., & Zapatka, E. (2017). Sensemaking in social media crisis communication – A case study on the Brussels bombings in 2016. *Proceedings of the 25th European Conference on Information Systems, 2169-2186* https://aisel.aisnet.org/ecis2017_rp/138

Mitra, T., & Gilbert, E. (2021). CREDBANK: A large-scale social media corpus with associated credibility annotations. *Proceedings of the International AAAI Conference on Web and Social Media, 9*(1), 258-267. https://doi.org/10.1609/icwsm.v9i1.14625

Mouratidis, D., Kanavos, A., & Kermanidis, K. (2025). From misinformation to insight: Machine learning strategies for fake news detection. *Information, 16*(3), 189. https://doi.org/10.3390/info16030189

Morris, D.S., & Morris, J.S. (2023). New social media nones: How and why Americans have changed their use of social media to consume political news. *Journal of Information, Communication and Ethics in Society, 21*(4), 468–484, https://doi.org/10.1108/JICES-04-2023-0052

Moran, B.B. & Morner, C. (2018). *Library and information center management.* Santa Barbara, California. Libraries Unlimited.

Mosallanezhad, A., Karami, M., Shu, K., Mancenido, M. V., & Liu, H. (2022). Domain adaptive fake news detection via reinforcement learning. *In Proceedings of the ACM Web Conference 2022*, *France*. https://doi.org/10.1145/3485447.3512258

Nakamura, K., Levy, S., & Wang, W. Y. (2020, May). Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *Proceedings of the Twelfth Language Resources and Evaluation Conference, France,* 6149-6157. https://aclanthology.org/2020.lrec-1.755/

Napoli, P. M. (2019). Algorithmic gatekeeping and the transformation of news organizations. *In Social media and the public interest: Media regulation in the disinformation age*. Columbia University Press. https://doi.org/10.7312/napo18454-004

Neely, S., Eldredge, C., & Sanders, R. (2021) Health information seeking behaviors on social media during the COVID-19 pandemic among American social networking site users: Survey study. *Journal of Medical Internet Research Publications, 23*(1). https://doi.org/10.2196/29802

Norman, D. A. (1988). *The psychology of everyday things*. Basic Books.

Oh, O., Eom, C., & Rao, H.R. (2015). Role of social media in social change: An analysis of collective sense making during the 2011 Egypt revolution. *Information Systems Research, 26*(1), 1-241. https://doi.org/10.1287/isre.2015.0565

Özçelik, O., Yenicesu, A.S., Yildirim, O., Haliloglu, D.S., Eroglu, E.E., & Can, F. (2023). *Cross-lingual transfer learning for misinformation detection in low-resource languages.* [PDF]. https://aclanthology.org/2023.ldk-1.59.pdf

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*(1–7), 107–117.

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you.* Penguin Press.

Park, S., & Lee, J. Y. (2023). Incidental news exposure on Facebook and its relation to trust in news. *Social Media + Society, 9*(1). https://doi.org/10.1177/20563051231158823

Pelrine, K., Imouza, A., Thibault, C., Reksoprodjo, M., Gupta, C., Christoph, J., & Rabbany, R. (2023). Towards reliable misinformation mitigation: Generalization, uncertainty, and gpt-4. *arXiv preprint.* https://doi.org/10.48550/arXiv.2305.14928

Pentina, I., & Tarafdar, M. (2014). From "information" to "knowing": Exploring the role of social media in contemporary news consumption. *Computers in Human Behaviour, 35*, 211-223. https://doi.org/10.1016/j.chb.2014.02.045

Perez, S. (2017, September 26). *Twitter trials expanding tweets from 140 characters to 280.* TechCrunch. https://techcrunch.com/2017/09/26/twitter-trials-an-expansion-beyond-140-characters/

Pérez-Escolar, M., Lilleker, D., & Tapia-Frade, A. (2023). A systematic literature review of the phenomenon of disinformation and misinformation. *Media and Communication*, *11*(2), 76–87. https://doi.org/10.17645/mac.v11i2.6453

Pew Research Center. (2025). *Social Media and News Fact Sheet.* https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/

Prybutok, G., & Ryan, S. (2015). Social media: the key to health information access for 18- to 30-year-old college students. *Computers, informatics, nursing, 33*(4), 132–141. https://doi.org/10.1097/CIN.0000000000000147

Qin, S., & Zhang, M. (2024). Boosting generalization of fine-tuning BERT for fake news detection. *Information Processing & Management, 61*(4), 103745. https://doi.org/10.1016/j.ipm.2024.103745

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1–67. https://doi.org/10.48550/arXiv.1910.10683

R Core Team. (2024). *R: A language and environment for statistical computing (Version 4.5.1)* [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

Reshi, J. A., & Ali, R. (2022). Online fake news detection using pre-trained embeddings. *2022 5th International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT),* 1–5. IEEE. https://doi.org/10.1109/IMPACT55510.2022.10029000

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533–536. https://doi.org/10.1038/323533a0

Russell, V. L., Lenth, R. V., Buerkner, P., Herve, M., & Love, J. (2025). *emmeans: Estimated marginal means, aka least-squares means* (R package). https://CRAN.R-project.org/package=emmeans

Sap, M., Card, D., Saadia Gabriel, S., Yejin Choi, Y., & Smith, N.A. (2019) The risk of racial bias in hate speech detection. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Italy,* 1668–1678. https://doi.org/10.18653/v1/P19-1163

Schafer, W.A., Ganoe, C.H., & Carroll, J.M. (2007). Supporting community emergency management planning through geocollaboration software architecture. *Computer Supportive Cooperative Work, 16,* 501-537

Shaeri, P., & Katanforoush, A. (2023). A semi-supervised fake news detection using sentiment encoding and LSTM with self-attention. *Proceedings of the International eConference on Computer and Knowledge Engineering*, 590–595. https://doi.org/10.1109/ICCKE60553.2023.10326287

Shanahan, M., McDonell, K. & Reynolds, L. (2023). Role play with large language models. *Nature 623*, 493–498. https://doi.org/10.1038/s41586-023-06647-8

Shannon, C. E., & American Telephone and Telegraph Company. (1948). *A mathematical theory of communication.* American Telephone and Telegraph Company.

Sheng, Q., Zhang, X., Cao, J., & Zhong, L. (2021). Integrating pattern and fact-based fake news detection via model preference learning. *Proceedings of the 30th ACM International*

*Conference on Information and Knowledge Management, Australia,* 1640-1650. https://doi.org/10.1145/3459637.3482440

Sheng, Q., Cao, J., Zhang, X., Li, R., Wang, D., & Zhu, Y. (2022). Zoom out and observe: News environment perception for fake news detection. *In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Ireland.* 4543-4556. https://doi.org/10.18653/v1/2022.acl-long.311

Shi, J., Zhao, X., Zhang, N., Lei, Y., & Min, L. (2023). Rough-Fuzzy Graph Learning Domain Adaptation for Fake News Detection. *Transactions on Computational Social Systems*, 11(5275-5286). https://doi.org/10.1109/TCSS.2023.3312182

Shin, Y., Sojdehei, Y., Zheng, L., & Blanchard, B. (2023). Content-based unsupervised fake news detection on Ukraine-Russia war. *SMU Data Science Review, 7*(3). https://scholar.smu.edu/datasciencereview/vol7/iss1/3

Shoemaker, Pamela J.; Vos, Tim P. (2009). *Gatekeeping Theory.* New York: Routledge. ISBN 978-0415981392.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter, 19*(1), 22-36. https://doi.org/10.1145/3137597.3137600

Sia, S., Dalima, A., & Mielke, S. J. (2020). Tired of topic models? Clusters of pretrained word embeddings make for fast and good topics too! *arXiv preprint.* arXiv:2004.14914

Silva, A., Luo, L., Karunasekera, S., & Leckie, C. (2021). Embracing domain differences in Fake news: Cross-domain fake News detection using multi-modal data. *Proceedings of the*

*AAAI Conference on Artificial Intelligence, 35*(1), 557-565. https://doi.org/10.1609/aaai.v35i1.16134

Sivakumar, M., Parthasarathy, S., & Padmapriya, T. (2024). Trade-off between training and testing ratio in machine learning for medical image processing. PeerJ. *Computer science, 10*, e2245. https://doi.org/10.7717/peerj-cs.2245

Smeros, P., Castillo, C., & Aberer, K. (2021). SciClops: Detecting and contextualizing scientific claims for assisting manual fact-checking. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, *Australia,* 1692–1702. https://doi.org/10.1145/3459637.3482475

Spalenza, M. A., Lusquino-Filho, L., França, F. M. G., Lima, P. M. V., & de Oliveira, E. (2021). Fake news detection using named entity recognition and part-of-speech sequences. CEUR *Workshop Proceedings, 2943*. https://ceur-ws.org/Vol-2943/fakedes_paper7.pdf

Spatharioti, S. E., Rothschild, D., Goldstein, D. G., & Hofman, J. M. (2025). Effects of LLM-based search on decision making: Speed, accuracy, and overreliance. *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems, Japan,* 1–15. https://doi.org/10.1145/3706598.3714082

Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S., & De Alfaro, L. (2017). Some like it hoax: Automated fake news detection in social networks. *arXiv preprint* arXiv:1704.07506

Tan, J., Yang, J., Wu, S., Chen, G., & Zhao, J. (2021). A critical look at the current train/test split in machine learning. *arXiv preprint*. arXiv:2106.04525.

Thackery, R., Crookston, B.T., West, J.H. (2013). Correlates of health-related social media use among adults. *Journal of Medical Internet Research Publications, 15*(1). https://doi.org/10.2196/jmir.2297

Tufekci, Z. (2015). Algorithmic harms beyond Facebook and Google: Emergent challenges of computational agency. *Colorado Technology Law Journal, 203.* https://scholar.law.colorado.edu/ctlj/vol13/iss2/4

UCLA: Statistical Consulting Group. (n.d.). *Introduction to generalized linear mixed models.* https://stats.oarc.ucla.edu/other/mult-pkg/introduction-to-generalized-linear-mixed-models/

UNESCO. (2023). *Survey on the impact of online disinformation and hate speech.* https://www.unesco.org/sites/default/files/medias/fichiers/2023/11/unesco_ipsos_survey.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*. https://doi.org/10.48550/arXiv.1706.03762

Vergho, T., Godbout, J. F., Rabbany, R., & Pelrine, K. (2024). Comparing GPT-4 and open-source language models in misinformation mitigation. *arXiv* preprint, abs/2401.06920.

Wang, Y., McKee, M., Torbica, A., & Stuckler, D. (2019). Systematic Literature Review on the Spread of Health-related Misinformation on social media. *Social Science & Medicine. 240*, 112552. https://doi.org/10.1016/j.socscimed.2019.112552

Vos, S.C., & Buckner, M.M. (2015). Social media messages in an emerging health crisis: Tweeting bird flu. *Journal of Health Communication, 21*(3), 301-308. https://doi.org/10.1080/10810730.2015.1064495

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380). 1146-1151. https://doi.org/10.1126/science.aap9559

Vrigazova, B. (2021). The proportion for splitting data into training and test set for the bootstrap in classification problems. *Business Systems Research: International Journal of the Society for Advancing Innovation and Research in Economy*, *12*(1), 228-242. https://doi.org/10.2478/bsrj-2021-0015

Wang, Y., Ma, F., Jin, Z., Yuan., Y., Zun, G., Jha, K., Su, L., & Gao, J. (2018). EANN: Event adversarial neural networks for multi-modal fake news detection. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, *United Kingdom,* 849–857. https://doi.org/10.1145/3219819.3219903

Wang, W.Y. (2017). Liar, liar pants on fire: A new benchmark dataset for fake news detection. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics Association for Computational Linguistics, Canada, 2,* 422–426. https://doi.org/10.18653/v1/P17-2067

Wardle, C., & Darakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe.*

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. (2022). Emergent abilities of large language models.

*Transactions on Machine Learning Research, 35.* https://doi.org/10.48550/arXiv.2206.07682

White, D. M. (1950). The "Gate Keeper": A case study in the selection of news. *Journalism Quarterly, 27*(4), 383-390. https://doi.org/10.1177/107769905002700403

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., & Yutani, H. (2023). *tidyverse: Easily install and load the tidyverse* (R package). https://CRAN.R-project.org/package=tidyverse

Wilson, T.D. (1981). On user studies and information needs. *Journal of Documentation, 37*(1), 3-15. https://doi.org/10.1108/eb026702

World Economic Forum. (2024). *The Global Risks Report 2024.* 18. https://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2024.pdf

Wu, L., Carley, K. M., & Liu, H. (2019). Misinformation in social media: Definition, manipulation, and detection. *SIGKDD Explorations, 21*(2), 80-90. https://doi.org/10.1145/3373464.3373475

Wu, Y., Wu, W., Xing, C., Zhou, M., & Li, Z. (2016). Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. *arXiv preprint* arXiv:1612.01627. https://doi.org/10.48550/arXiv.1612.01627

Wu, Y., Li, Z., Wang, C., Zheng, H., Zhao, S., Li, B., & Tao, D. (2024). Domain re-modulation for few-shot generative domain adaptation. *Advances in Neural Information Processing Systems, 36*. https://doi.org/10.48550/arXiv.2302.02550

X. (January 2025). *Help Center – Authenticity.* https://help.x.com/en/rules-and-policies/authenticity

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems, 32*. https://doi.org/10.48550/arXiv.1906.08237

Yang, J., Hui, B., Yang, M., Wang, B., Li, B., Li, F., & Li, Y. (2024). Iterative forward tuning boosts in-context learning in language models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, *Thailand, (Volume 1: Long Papers),* 15460–15473. https://doi.org/10.48550/arXiv.2305.13016

Yin, S., Zhu, P., Wu, L., Gao, C., & Wang, Z. GAMC: An unsupervised method for fake news detection using graph autoencoder with masking. *Proceedings of the AAAI Conference on Artificial Intelligence, Canada, 38*(1), 347–355. https://doi.org/10.1609/aaai.v38i1.27788

Zhang, X., Cao., J., Li., X., Sheng, Q., Zhong, L., & Shu, K. (2021). Mining dual emotion for fake news detection. *Proceedings of the Web Conference 2021, USA,* 3465-3476. https://doi.org/10.1145/3442381.3450004

Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., Li, J., Hu, R., Zhang, T., Wu, F. (2023). Instruction tuning for large language models: A survey. *arXiv preprint.* arXiv:2308.10792.

Zhou, S. (2025). Effectiveness of citizen journalism: evaluating the credibility, bias, and impact of user-generated news content in the digital age. *Humanities & Social Sciences Communications, 12*(1956). https://doi.org/10.1057/s41599-025-06216-6

Zugecova, A., Macko, D., Srba, I., Moro, R., Kopál, J., Marcinčinová, K., & Mesarčík, M. (2025). Evaluation of LLM vulnerabilities to being misused for personalized disinformation generation. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, Austria, (Volume 1: Long Papers),* 780-797. https://doi.org/10.18653/v1/2025.acl-long.38

# Appendices

## Appendix A: Full prompt specifications

### Zero-shot prompt specification

We used zero-shot prompting to evaluate LLM performance in this study's misinformation detection task. In this condition, the model was provided with only a short instruction and the claim to be classified. No labeled examples were included in the prompt. As in the few-shot condition, the model was instructed to respond with a single-word output indicating the predicted label.

**Figure 10**

*Zero-Shot Prompt Template*

```
35
36      def generate_prompt(self, claim, domain):
37          return (
38              f"Classify the following {domain} claim from {structure} content as 'true' or 'false'. "
39              f"Respond with only one word (true/false).\n\n"
40              f"Claim: \"{claim}\"\nAnswer:"
41          )
```

The placeholders were dynamically populated using values from the dataset. The {domain} variable corresponded to the topical category of the claim (Health, Politics, or War), while {structure} indicated whether the claim originated from UGC or NGC. Unlike the few-shot condition, the prompt did not include labeled training examples, meaning that classification decisions relied entirely on the model's pre-trained knowledge rather than contextual examples.

### Generation parameters

Model inference was conducted using each of this study's sampled LLMs. The generation parameters were configured to produce simple, short responses appropriate for a binary classification task.

We chose a temperature of 0.1 to minimize randomness in token selection and encourage the model to assign the most probable label. Although we tested temperatures between 0.0 and 0.7, we found that 0.1 produced the most consistent and favourable results across the sampled LLMs.

The maximum generation length was limited to 5 tokens, ensuring concise responses while allowing minor formatting variations. This constraint was also a factor in the decision not to evaluate reasoning variants of the sampled LLMs. We found that when reasoning functions were enabled, models disregarded prompt specifications and responded with a single word, instead generating long explanations before producing a final answer. This behaviour substantially increased the required token generation length, introducing additional variability in response formatting and increasing computational demands during local testing. Limiting token length ensured consistent single-word classification outputs and supported efficient evaluation within the available computational environment.

To ensure methodological rigor, the experiment was repeated across five runs using different random seeds (42, 123, 456, 789, 999). A random seed is a number used to make random processes in a computer repeatable. By setting a specific seed value, the same sequence of "random" choices can be reproduced whenever the experiment is run again. In the zero-shot condition, the seed controlled minor variation in the model's token generation process. Using

different seeds ensured that results were not dependent on a single random sampling configuration while maintaining full reproducibility.

**Few-shot prompt specification**

We used few-shot prompting to evaluate LLM performance in this study's misinformation detection task. Each prompt consisted of a short instruction, followed by several labeled examples drawn from the training data, and then the target claim to be classified. The model was instructed to respond with a single-word output indicating the predicted label.

**Figure 11**

*Few-Shot Prompt Template*

```
    def generate_prompt(self, claim, domain, structure, few_shot_examples):
        """
        Generate prompt with few-shot examples from training data.
        """
        prompt = (
            f"Classify the following {domain} claim from {structure} content as true or false. "
            f"Respond with ONLY one word: true or false.\n\n"
            f"Examples:\n"
        )
        for idx, row in few_shot_examples.iterrows():
            example_claim = row['claim']
            example_structure = row.get('structure', 'user')
            example_label = 'true' if row['label'] == 1 else 'false'

            prompt += f"Claim: \"{example_claim}\"\n"
            prompt += f"Structure: {example_structure}\n"
            prompt += f"Answer: {example_label}\n\n"

        prompt += f"Now classify the following {domain} claim from {structure} content.\n\n"
        prompt += f"Claim: \"{claim}\"\n"
        prompt += f"Answer:"

        return prompt
```

The placeholders were dynamically populated using values from the dataset. The {domain} variable corresponded to the topical category of the claim (Health, Politics, or War), while {structure} indicated whether the claim originated from UGC or NGC.

**Few-Shot example selection**

For each experimental run, eight few-shot examples were sampled from the training portion of the dataset. Examples were selected using a stratified sampling strategy, which ensured representation across both domain and class label. Specifically, examples were grouped by domain (Health, Politics, War), label (true or false), and sub-dataset (NGC or UGC).

**Generation parameters**

Model inference was conducted using each of this study's sampled LLMs. The generation parameters were configured to produce simple, short responses appropriate for a binary classification task. We chose a temperature of 0.1, a maximum token length of 5, and newline as the top sequence function for the same reasons described in the zero-shot prompt specification section.

The few-shot condition was repeated across five runs using different random seeds (42, 123, 456, 789, 999). In this condition, using different seeds across runs ensured that results were not dependent on a model's token or a single random selection of examples. Instead, each seed produced minor randomness in the model's token generation process and a different set of few-shot examples, allowing the prompt examples to vary across runs while maintaining full reproducibility.

## Appendix B: GLMMs

Appendix D presents the full outputs of the GLMMs used to evaluate the effects of content type, domain, prompting strategy, and model architecture on classification accuracy.

Models were estimated using logistic regression with random intercepts for both claim identifier and model identifier.

**Table 7**

GLMM: Content type × Domain × Prompting

| Predictor | β | SE | z | p |
|---|---|---|---|---|
| Intercept | 2.128 | 0.229 | 9.270 | < .001 |
| Content Type (UGC) | -0.699 | 0.267 | -2.61 | 0.008 |
| Domain | -0.218 | 0.089 | -2.44 | 0.014 |
| Prompting (Zero-Shot) | -0.876 | 0.169 | -5.188 | < .001 |
| Content Type × Domain | 0.178 | 0.122 | 1.450 | .0.146 |
| Content Type × Prompting | 0.520 | 0.227 | 2.290 | 0.021 |
| Domain × Prompting | 0.176 | 0.077 | 2.276 | 0.022 |
| Content Type × Domain × Prompting | -0.242 | 0.104 | -2.321 | 0.020 |

*Note*. The model was estimated using maximum likelihood with Laplace approximation. Random intercepts were included for model identifier and claim identifier to account for repeated observations across models and items. The dependent variable was binary classification accuracy (0 = incorrect, 1 = correct). Reference categories were NGC, the Health domain, and Few-shot prompting. Coefficients represent changes in the log-odds of a correct classification. Positive coefficients indicate increased odds of correct classification relative to the reference condition.

**Model Fit Statistics**

AIC = 14,437.60

BIC = 14,512.90

Log Likelihood = -7,208.80

Number of observations = 13,874

Residual degrees of freedom = 13,864

**Table 8**

GLMM: Content type × Domain × Architecture

| Predictor | β | SE | z | p |
|---|---|---|---|---|
| Intercept | 1.013 | 0.104 | 9.717 | < .001 |
| Content Type (UGC) | -0.233 | 0.144 | -1.611 | 0.107 |
| Domain2 (Politics) | -0.205 | 0.146 | -1.40 | 0.161 |
| Architecture (BERT) | 0.705 | 0.240 | 2.936 | 0.003 |
| Content Type × Domain2 | 0.032 | 0.205 | 0.156 | 0.875 |
| Content Type × Domain3 (War) | -0.144 | 0.203 | -0.709 | 0.478 |
| Content Type × Architecture | -0.235 | 0.326 | -0.721 | 0.470 |
| Domain2 × Architecture | -0.425 | 0.319 | -1.331 | 0.183 |
| Domain3 × Architecture | -0.008 | 0.336 | -0.027 | 0.978 |
| Content Type × Domain2 × Architecture | 0.811 | 0.454 | 1.785 | 0.074 |

| Predictor | β | SE | z | p |
|---|---|---|---|---|
| Content Type × Domain3 × Architecture | -0.011 | 0.455 | -0.026 | 0.979 |

*Note*. The model was estimated using maximum likelihood with Laplace approximation. The dependent variable was binary classification accuracy (0 = incorrect, 1 = correct). Reference categories were NGC, the Health domain, and Architecture. Coefficients are reported in log-odds. Positive coefficients represent changes in the log-odds of a correct classification.

**Model Fit Statistics**

AIC = 8309.7

BIC = 8398.6

Log Likelihood = -4141.9

Number of observations = 6883

Residual degrees of freedom = 6870

## Appendix C: Descriptive results

Appendix C presents descriptive performance statistics for all evaluated models across domains and content types. Results are reported separately for the UGC and NGC datasets. For each model and domain, we report accuracy (Acc), precision (Pre), and F1-score (F1). These descriptive statistics provide a detailed breakdown of model performance prior to the inferential statistical analysis reported in Appendix B.

Tables 8 and 9 report mean performance across five experimental runs for the UGC and NGC datasets, respectively. Table 10 reports the change in F1-score between zero-shot (ZS) and

few-shot (FS) prompting conditions across both datasets, and Figure 12 illustrates the change in mean F1-score between prompting conditions.

**Table 9**

*Domain-Specific Results in the UGC Dataset*

| | Health | | | Politics | | | War | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | Acc | Pre | F1 | Acc | Pre | F1 | Acc | Pre | F1 |
| BERT | 0.653 | 0.625 | 0.689 | 0.745 | 0.882 | 0.697 | 0.673 | 0.695 | 0.653 |
| RoBERTa | 0.750 | 0.696 | 0.779 | 0.823 | 0.904 | 0.808 | 0.788 | 0.826 | 0.775 |
| DistilBERT | 0.653 | 0.625 | 0.689 | 0.823 | 0.947 | 0.800 | 0.634 | 0.612 | 0.666 |
| Gemma3-4b-zs | 0.542 | 0.399 | 0.361 | 0.528 | 0.267 | 0.346 | 0.483 | 0.242 | 0.326 |
| Gemma3-27b-zs | 0.784 | 0.811 | 0.779 | 0.78 | 0.834 | 0.772 | 0.730 | 0.754 | 0.724 |
| Llama3-4b-zs | 0.634 | 0.753 | 0.586 | 0.525 | 0.754 | 0.401 | 0.603 | 0.712 | 0.545 |
| Llama3-70b-zs | 0.750 | 0.777 | 0.743 | 0.717 | 0.806 | 0.698 | 0.665 | 0.735 | 0.638 |
| Mistral7b-zs | 0.719 | 0.746 | 0.711 | 0.686 | 0.766 | 0.664 | 0.676 | 0.738 | 0.675 |
| Mixtral8x7b-zs | 0.692 | 0.710 | 0.685 | 0.784 | 0.827 | 0.778 | 0.765 | 0.791 | 0.759 |
| Gemma3-4b-fs | 0.606 | 0.659 | 0.559 | 0.806 | 0.824 | 0.801 | 0.715 | 0.784 | 0.695 |
| Gemma3-27b-fs | 0.800 | 0.800 | 0.800 | 0.858 | 0.865 | 0.858 | 0.788 | 0.790 | 0.788 |
| Llama3-4b-fs | 0.619 | 0.618 | 0.617 | 0.595 | 0.625 | 0.556 | 0.563 | 0.572 | 0.558 |
| Llama3-70b-fs | 0.773 | 0.773 | 0.772 | 0.866 | 0.889 | 0.865 | 0.719 | 0.720 | 0.718 |
| Mistral7b-fs | 0.745 | 0.750 | 0.744 | 0.806 | 0.807 | 0.806 | 0.758 | 0.759 | 0.758 |
| Mixtral8x7b-fs | 0.761 | 0.753 | 0.754 | 0.833 | 0.831 | 0.827 | 0.752 | 0.754 | 0.751 |

*Note.* Results represent mean performance across five experimental runs. Metrics are reported separately for each topical domain within the UGC dataset. “ZS” denotes zero-shot prompting and “FS” denotes few-shot prompting.

**Table 10**

*Domain-Specific Results in the NGC Dataset*

| | Health | | | Politics | | | War | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | Acc | Pre | F1 | Acc | Pre | F1 | Acc | Pre | F1 |
| BERT | 0.788 | 0.727 | 0.814 | 0.635 | 0.613 | 0.667 | 0.731 | 0.714 | 0.741 |
| RoBERTa | 0.692 | 0.632 | 0.750 | 0.827 | 0.758 | 0.847 | 0.904 | 0.889 | 0.906 |
| DistilBERT | 0.788 | 0.727 | 0.814 | 0.654 | 0.667 | 0.64 | 0.788 | 0.800 | 0.784 |
| Gemma3-4b-zs | 0.461 | 0.230 | 0.315 | 0.462 | 0.231 | 0.316 | 0.498 | 0.249 | 0.332 |
| Gemma3-27b-zs | 0.883 | 0.883 | 0.883 | 0.796 | 0.822 | 0.791 | 0.817 | 0.821 | 0.817 |
| Llama3-4b-zs | 0.712 | 0.749 | 0.702 | 0.661 | 0.752 | 0.628 | 0.67 | 0.704 | 0.654 |
| Llama3-70b-zs | 0.821 | 0.821 | 0.821 | 0.846 | 0.854 | 0.845 | 0.775 | 0.775 | 0.775 |
| Mistral7b-zs | 0.805 | 0.814 | 0.804 | 0.738 | 0.772 | 0.738 | 0.810 | 0.820 | 0.808 |
| Mixtral8x7b-zs | 0.832 | 0.852 | 0.83 | 0.773 | 0.818 | 0.764 | 0.794 | 0.818 | 0.79 |
| Gemma3-4b-fs | 0.787 | 0.807 | 0.780 | 0.696 | 0.698 | 0.695 | 0.750 | 0.780 | 0.743 |
| Gemma3-27b-fs | 0.875 | 0.875 | 0.875 | 0.767 | 0.747 | 0.751 | 0.782 | 0.783 | 0.782 |
| Llama3-4b-fs | 0.771 | 0.774 | 0.770 | 0.725 | 0.742 | 0.722 | 0.714 | 0.714 | 0.713 |
| Llama3-70b-fs | 0.805 | 0.806 | 0.805 | 0.873 | 0.88 | 0.865 | 0.821 | 0.804 | 0.865 |
| Mistral7b-fs | 0.864 | 0.816 | 0.864 | 0.805 | 0.816 | 0.803 | 0.816 | 0.820 | 0.815 |
| Mixtral8x7b-fs | 0.853 | 0.863 | 0.851 | 0.752 | 0.762 | 0.750 | 0.760 | 0.761 | 0.759 |

*Note.* Results represent mean performance across five experimental runs. Metrics are reported separately for each topical domain within the NGC dataset.

**Table 11**

*Few-Shot Performance Gain by Content Type*

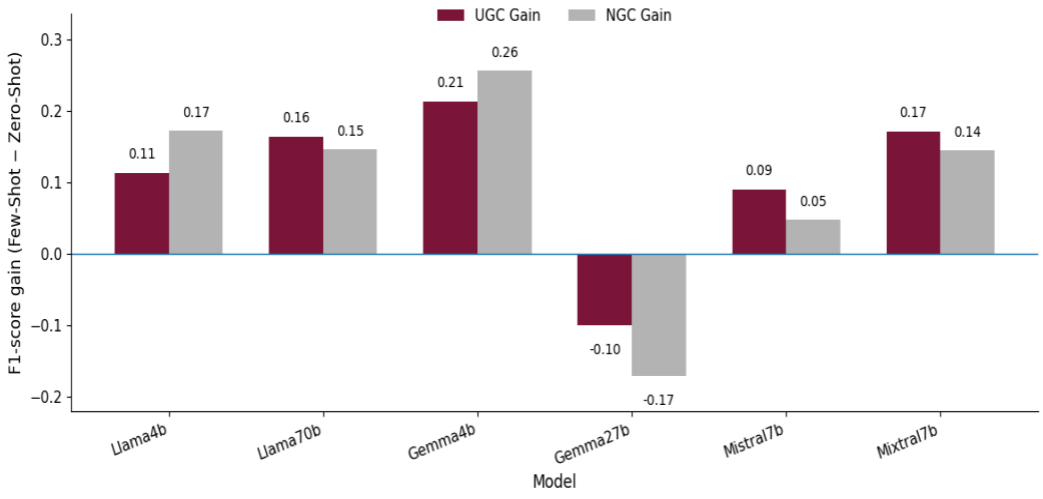
| Model | UGC-ZS | UGC-FS | +/- from ZS | NGC-ZS | NGC-FS | +/- from ZS |
|---|---|---|---|---|---|---|

| | | | | | | |
|---|---|---|---|---|---|---|
| Gemma3-4b | 0.360 | 0.573 | 0.213 | 0.344 | 0.600 | 0.256 |
| Gemma3-27b | 0.754 | 0.654 | - 0.100 | 0.825 | 0.654 | - 0.171 |
| Llama3-4b | 0.468 | 0.581 | 0.113 | 0.557 | 0.729 | 0.172 |
| Llama3-70b | 0.619 | 0.783 | 0.164 | 0.662 | 0.808 | 0.146 |
| Mistral7b | 0.678 | 0.768 | 0.090 | 0.781 | 0.828 | 0.047 |
| Mixtral8x7b | 0.608 | 0.779 | 0.171 | 0.640 | 0.784 | 0.144 |

*Note.* Positive values indicate improved performance when few-shot examples were provided, while negative values indicate a decrease in performance relative to the zero-shot baseline.

**Figure 12**

*Few-Shot Performance Gain by Model and Content Type*




*Note.* Metrics shown are calculated from mean overall performance. F1-scores below the zero baseline denote a decrease in F1-scores from the zero-shot condition. F1-scores above the zero-baseline denote an increase in f1-scores from the zero-shot condition.