

Analyzing Artificial Intelligence's Ability to Detect Misinformation

By

Maxwell Bilyk, B.S Data Science

A thesis submitted to the Graduate Committee of
Ramapo College of New Jersey in partial fulfillment
of the requirements for the degree of
Master of Science in Data Science
Spring, 2025

Committee Members:

Kathleen Walsh, Advisor

Amanda Beecher, Reader

Kelly Fitzpatrick, Reader

COPYRIGHT

© Maxwell Bilyk

Table of Contents	4
Abstract	1
Chapter 1: Introduction	3
Chapter 2: Understanding Misinformation	5
Chapter 3: Existing Solutions to Tackle Misinformation	10
Chapter 4: AI and its ability to detect misinformation	13
Chapter 5: Case Studies	18
Chapter 6: Ethical Implications of AI Misinformation Detection	23
Chapter 7: Challenges in Implementing AI Solutions	27
Chapter 8: Performance Analysis	32
Methodology	32
Results and Interpretation	33
Implications	35
Limitations	36
Conclusion	36
Chapter 9: Future Directions and Recommendations	37
Development of Multimodal Misinformation Detection Systems	37
Incorporation of Explainable AI (XAI) Frameworks	38
Ethical AI Design and Bias Mitigation	40
AI Literacy and Public Education Programs	43
Human-in-the-Loop Systems for High-Stakes Domains	45
Conclusion	48
Chapter 10: Conclusion	50
References	55
Appendix	57

Abstract

Misinformation and disinformation represent critical societal challenges of the 21st century, significantly amplified by rapid advancements in digital technology. The proliferation of generative artificial intelligence (AI) exacerbates these problems, enabling false narratives to spread at unprecedented speeds, undermining public trust, polarizing societies, and endangering democratic processes. Traditional methods, such as manual fact-checking, governmental initiatives, and educational programs, while effective, are increasingly insufficient in addressing the scale and immediacy of digital misinformation.

This thesis aims to critically evaluate artificial intelligence's potential in addressing the misinformation crisis. Specifically, it investigates how AI-driven techniques, particularly natural language processing (NLP), can improve misinformation detection and fact-checking processes. Further, it examines ethical considerations surrounding AI use, evaluates practical and technical implementation challenges, and proposes solutions to improve these technologies.

A mixed-methods approach was employed, encompassing historical analysis of misinformation, review of existing solutions, examination of contemporary AI technologies, and detailed case studies evaluating AI's application in real-world misinformation scenarios. Additionally, a quantitative performance analysis of an AI-driven misinformation classifier was conducted using a structured prompt engineering method. This involved scoring news articles on factuality, logic, sentiment, and bias, using a composite measure tested against a labeled dataset of verified true or false articles.

The thesis demonstrated that AI systems, particularly large language models (LLMs), show substantial promise in misinformation detection, achieving over 90% accuracy when optimally calibrated. Real-world case studies, including the UK-based organization Full Fact,

revealed AI's capacity to significantly enhance fact-checking efficiency and responsiveness. Nevertheless, the study identified critical limitations, including AI's difficulties in nuanced contextual understanding, bias propagation, ethical dilemmas, and environmental sustainability concerns. The research highlights the necessity of continued human oversight—particularly through human-in-the-loop (HITL) models—to address AI's current limitations.

The findings underscore that AI, while not flawless, holds promise as a scalable, effective tool against misinformation when complemented by rigorous ethical frameworks, transparency (via Explainable AI), multimodal approaches, human-in-the-loop systems and widespread AI literacy initiatives. The broader implications suggest that successful deployment of AI in misinformation detection necessitates interdisciplinary collaboration, proactive bias mitigation, robust public education, and sustained human involvement. Addressing misinformation through AI is not only a technological pursuit but fundamentally an ethical and societal responsibility crucial for maintaining the integrity of public discourse and democratic institutions in the digital age.

Chapter 1: Introduction

Misinformation and disinformation are among the most pressing issues of the 21st century. Misinformation is when information is false, but not created with the intention of causing harm, while disinformation is when the information presented is false and was deliberately made to be false to harm a person or group (Commons Library 2024). Whether the information is falsified on purpose or not, the underlying consequences of false information remain the same. The rise of digital media has enabled the rapid spread of information, at speeds previously unimaginable. Governments and individuals alike have leveraged this capability to disseminate harmful and false narratives. For example, the Russian government's systematic use of disinformation, including targeting religious minorities, demonstrates the global reach and harmful potential of false narratives. A problem exacerbated by advancements in digital technology, including generative AI. Russia weaponizes antisemitism and Holocaust distortion to shape international discourse on its war in Ukraine. When Russian President Vladimir Putin announced Russia's "special military operation" in Ukraine, he claimed the goal was to "denazify" the country, falsely accusing Ukrainian officials of being neo-Nazis and committing "genocide" against Russian-speaking citizens. Russian state media perpetuated this false narrative, often devolving into antisemitism and Holocaust distortion. (Weiner 2024)

Misinformation's detrimental effects are not limited to state actors. A prominent example involving an individual is Alex Jones, who propagated falsehoods about the Sandy Hook Elementary School shooting. On his show, InfoWars, he claimed the massacre—where 20 children and six educators were killed—was a hoax orchestrated by "crisis actors" to promote stricter gun control laws. These false claims led to severe harassment of grieving families, some

of whom were forced to relocate multiple times. Jones was ordered to pay nearly \$1.5 billion in damages, as his actions caused demonstrable harm. (Brodsky 2024)

These examples highlight the global scale of the misinformation crisis, a problem only exacerbated by the advent of generative AI. Misinformation is not a new phenomenon; it has existed for centuries. What is new, however, is the speed and ease with which false content can be created and shared in the digital age. With access to a platform and generative AI, an individual can craft an article of misinformation and distribute it to thousands within minutes. The speed of this process renders traditional fact-checking methods—once effective—unable to scale to the same magnitude as the spread of misinformation.

This raises a critical question: How can society combat misinformation in the digital era? The solution may lie in the same technological advancements that have fueled its proliferation. This thesis aims to critically evaluate the potential of artificial intelligence's role in combating misinformation, focusing on its technical capabilities in natural language processing (NLP), in political and health-related misinformation, and the ethical implications of its widespread adoption into society.

This thesis is organized into nine chapters, each addressing a key aspect of the research. Each chapter contributes to a layered understanding of the problem, starting with the historical and technical foundations of misinformation and progressing to actionable solutions and future directions for AI-driven interventions. The progression of chapters begins with defining the problem of misinformation and then progresses into actionable solutions using artificial intelligence.

Chapter 1 introduces the problem of misinformation, its significance, and the objectives of this study. Chapter 2 lays the foundation by defining misinformation and exploring its history

and mechanisms of spread. Chapter 3 reviews existing non-AI solutions, highlighting their strengths and limitations.

Chapter 4 delves into the role of AI, explaining its capabilities and current applications in misinformation detection. Chapter 5 presents real-world case studies that illustrate the successes and challenges of using AI in this domain. Chapter 6 critically examines ethical implications such as bias, accountability, and privacy concerns. Chapter 7 addresses practical challenges, including technical limitations, costs, and stakeholder resistance. Chapter 8 proposes future directions, offering actionable solutions and a long-term roadmap for ethical AI use in combating misinformation. Finally, Chapter 9 concludes the thesis by summarizing findings and reflecting on the broader implications of the research.

Chapter 2: Understanding Misinformation

When information is presented, it is typically assumed to be factual. However, when this assumption is incorrect, the incorrect information falls into one of three categories:

1. **Misinformation** is false, but not created or shared with the intention of causing harm.
2. **Disinformation** is deliberately created to mislead, harm, or manipulate a person, social group, organization, or country.
3. **Malinformation** is based on fact, but used out of context to mislead, harm, or manipulate. An example of malinformation is editing a video to remove important context to harm or mislead. (Cisa.gov)

These definitions, drawn from the Cybersecurity and Infrastructure Security Agency (CISA), highlight the nuanced and multifaceted nature of misinformation. This chapter explores the historical evolution of misinformation and examines its amplification in the digital era,

setting the stage for analyzing AI's potential role in identifying and addressing incorrect information.

Misinformation has a complex foundation and has negatively impacted society throughout history. While often viewed as a modern phenomenon, its roots extend back centuries. One of the earliest documented cases of disinformation occurred during the rule of Roman Emperor Augustus Octavian. To justify his instigation of a civil war, Octavian orchestrated a campaign of disinformation, spreading propaganda through word of mouth—the primary method of communication at the time (Sifuentes 2019). This demonstrates that even without advanced technology, humans have been able to find ways to spread misinformation in order to achieve political goals.

The invention of the printing press in the 15th century marked a significant milestone in the history of misinformation. For the first time, information could be disseminated rapidly and to a broad audience. It was much easier to distribute printed material to the masses and have them read it at their own time rather than the traditional spread of information through word of mouth. While this advancement improved access to knowledge, it also facilitated the spread of false information. For instance, printed pamphlets during the Protestant Reformation often contained exaggerated or outright false claims designed to sway public opinion. This era highlights the dual-edged nature of technological advancements in communication: they can empower both truth and falsehoods.

In 1782, Benjamin Franklin demonstrated the use of misinformation as a propaganda tool. He fabricated an issue of a Boston newspaper to garner sympathy for the American Revolution. This fake publication falsely claimed that British forces had hired Native Americans to commit atrocities against civilians, stoking public hostility toward both groups. The public

perception of Native Americans and the British Empire was forever tainted in the eyes of the colonists, which helped foster support for the American Revolution. Franklin's actions illustrate how misinformation has been strategically employed to manipulate public sentiment, with lasting societal consequences.

From the early days of Emperor Octavian in Rome all the way to Benjamin Franklin in America, it is obvious that misinformation has been a tool used to manipulate the public. As we step into the 20th century we see a rapid change in the reach of misinformation accelerated by advances in information technology. The introduction of radio in the 1930s created a new platform for both accurate reporting and the dissemination of falsehoods. Despite the emergence of radio, newspapers remained the primary source of information until the mid-20th century, ensuring that misinformation's reach was still somewhat limited by the slower pace of print media.

The advent of the internet in the 1990s marked a seismic shift in how information is shared and consumed. Personal computers, and later smartphones, enabled instantaneous communication and access to vast amounts of information. This digital revolution gave rise to new forms of media, including online news outlets and social media platforms, exponentially increasing the speed and scale at which information could be distributed.

This unprecedented connectivity has made misinformation a much larger problem compared to previous societies due to the ease of access brought forth by cellphones and social media. Social media platforms, in particular, have become breeding grounds for the rapid dissemination of false information with "An estimated 500,000 deepfakes were shared on social media in 2023" (Konopliov 2024). One key factor driving this phenomenon is the infrastructure of social media algorithms. These algorithms, often referred to as "black box models" in machine

learning terminology, are designed to maximize user engagement by curating content that aligns with individual preferences. However, this design inadvertently amplifies misinformation. Content that elicits strong emotional reactions, such as anger or fear, tends to perform well, leading algorithms to prioritize it in users' feeds. As a result, misinformation can go viral faster than factual corrections or clarifications, creating a feedback loop that perpetuates its spread (Pangotra 2024).

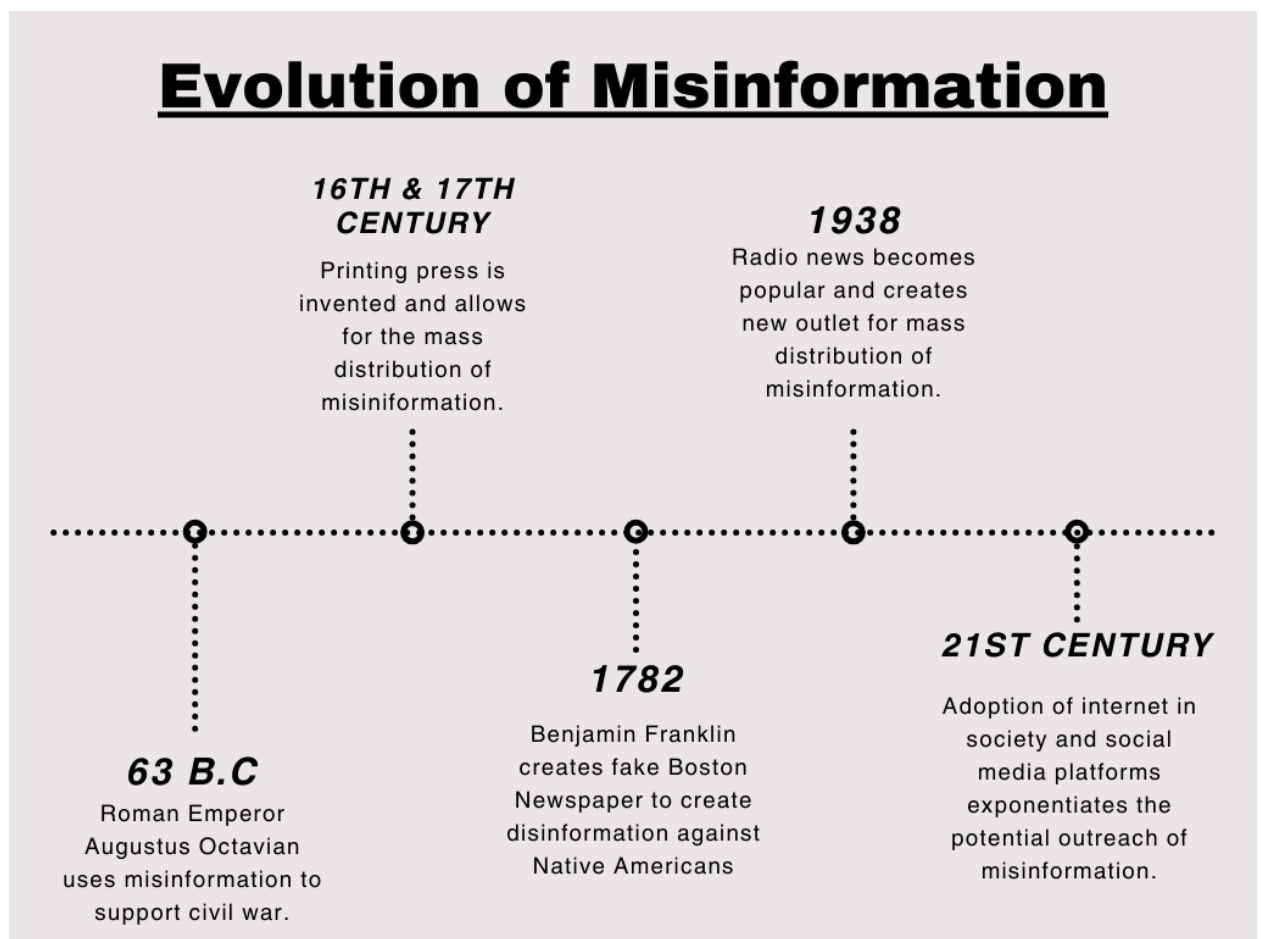


Figure 1. Timeline representing the evolution of misinformation throughout human history.

Highlighting examples of misinformation and other notable events related to misinformation

Misinformation in the digital age is not merely a technical challenge; it has profound social, political, and ethical implications. False information can undermine public trust in

institutions, polarize societies by reinforcing biases, and even influence elections. Moreover, the rapid proliferation of misinformation complicates efforts to correct it. The University of Western Australian conducted a study where psychologists asked students to read an account of an accident involving a busload of elderly people. Then the students were told that they were not elderly. For some the information ended there and for others they were told it was actually a busload of hockey players. The students were at a later date told to make inferences about the accident and the results were very interesting. Even the control group, that was told to be wary of misinformation, ended up agreeing with statements like “They struggled to exit the vehicle because they were frail”. This study highlights the lingering effects of misinformation, which makes a serious emphasis on getting a story straight before it hits a mass audience (Ross 2011).

The complexity of addressing misinformation in the digital age underscores the importance of exploring innovative solutions. While this thesis does not aim to provide strategies for distinguishing between misinformation, disinformation, and malinformation, it focuses on analyzing the potential for artificial intelligence to detect incorrect information. A critical yet unanswered question is whether AI can not only identify falsehoods but also discern their intent and context, thereby differentiating among the three categories of misinformation, which will be explored in Chapter 4.

From the word-of-mouth campaigns of ancient Rome to the algorithm-driven misinformation of the digital era, the evolution of misinformation reveals its persistent and adaptable nature. Each technological advancement in communication has brought both opportunities for enlightenment and challenges in combating falsehoods. As we continue to navigate the complexities of misinformation, the role of emerging technologies, particularly artificial intelligence, will be crucial in shaping the future of information integrity. This chapter

has set the foundation for examining AI's capabilities in addressing misinformation, providing historical context to better understand the scope of the problem. The subsequent chapters will look to explore how we have combated misinformation in the past—and why these methods may no longer be an effective solution to the growing misinformation crisis.

Chapter 3: Existing Solutions to Tackle Misinformation

In today's digital era, there is a wide range of ways fact checking or media literacy initiatives are pushed onto the public. There are manual approaches, policy efforts from governments and institutions, as well as platform based strategies all in place to try and limit the spread of misinformation. This chapter will look to examine these systems and discuss their benefits and highlight their limitations with the introduction of generative AI.

Manual fact-checking involves individuals or teams analyzing content by cross-referencing it with existing knowledge and research to assess its accuracy. There are currently some large fact-checking organizations that have large teams of fact checkers and publish curated content that has been approved by their teams. One of these organizations is factcheck.org. They are a verified signatory of the International Fact-Checking Network, where they have independent fact checkers working for them that are evaluated annually (FactCheck.org). Additionally, they do publish their own content which analyzes political events throughout the world and aims to give an unbiased breakdown of true and false claims made by an individual. The mission of this organization is to shed some light on the purposefully confusing world of U.S. politics.

Organizations like factcheck.org work with their teams to read through all forms of media and look for supporting or refuting evidence for these claims. This manual form of fact-checking

is an extremely accurate way to detect misinformation through rigorous research and comparison. This being said, there's more writers than there are fact checkers, so there will always be a large majority of content that society comes into contact with that has not been reviewed, which is a large reason why governments and institutions have made initiatives to promote media literacy amongst individuals.

Recognizing the limitations of manual fact-checking, government agencies have also taken steps to enhance public awareness and resilience against misinformation. The Department of Homeland Security has made efforts to promote media literacy and critical thinking online. They provide information for “how individuals can be resilient to harmful or false narratives in the online space and effective when consuming and sharing information online. The resource provides definitions of the various types of misleading information, offers recommendations for approaching digital content, and lists further educational and programmatic resources.” (Department of Homeland Security 2021). In addition to this government push, some institutions, like the National Association of State Boards of Education, have looked to bring media literacy into the education system in order to produce a younger generation of better informed people. This is a very interesting initiative as it will help students become aware of misinformation through their education and beyond (Alkam & DiGiacomo 2024).

Another significant area where misinformation spreads is social media platforms. Platforms like Facebook and Instagram were partnered with third party fact checking organizations. These organizations basically operated in their normal manual fact checking fashion, but they were specifically fact-checking claims from social media posts, and then those posts could be tagged as misinformation and give people a warning about potential misinformation within the post and a brief explanation why. Interestingly, Meta recently ended

these partnerships because of worries that these organizations were limiting freedom of expression (about.fb.com 2025). These programs still exist for other countries through Meta, just no longer in the US. While these systems were effective in a vacuum at first, once they were fully challenged with the scale of misinformation online and the addition of generative AI, these systems began to seriously underperform. Additionally, the inherent algorithmic bias in users' social media feeds also feeds into spreading misinformation. The goal of most of these algorithms is to keep users on the platform as long as possible, which can be dangerous because the algorithm may boost misinformation as long as it is keeping more people engaged with the app. This creates a conflict of interest between the fact checkers of the platform and the executives pushing for more screen time. There were also concerns over inconsistent enforcement of fact-checking on social media platforms. Many individuals believed only certain issues and people of certain political affiliations were being targeted. Mark Zuckerberg, CEO of Meta, was recorded on a podcast saying that the Biden administration encouraged them to remove posts and enforce fact-checking on issues related to the covid vaccine (Novet 2025). Although the initiative was likely meant to tackle health misinformation, many individuals did not see it in a positive light. This led to the eventual end of Meta conducted fact-checking on their platform because of the community response. This use case showcased how misinformation itself can be biased, as people are much more likely to believe in what they wish to be true, and will label the truth as “false news”.

Despite these ongoing efforts, the emergence of generative AI has introduced new challenges that threaten to outpace existing fact-checking methods. This technology has enabled instant creation of misinformation in both short and long term content. The spontaneous nature of this technology in tandem with the instant connectivity available through social media,

presents a challenge that modern solutions to misinformation may not be able to scale with. Manual fact checking is a very slow process to ensure that no stone is left unturned. This leads to the fact checking process to take days, weeks and even months (Baker & Fairbank 2022). A study from *Social Insider* found that an average instagram user gets 2635 impressions per post (Socialinsider.io). This means that any average user can present misinformation to thousands of people for every post that they make, and then each of those people can pass along the information to their networks and so on. As previously mentioned, studies show that even after misinformation has been fact-checked, there are diminishing effects on people's understanding of the story.

As social media and generative AI continue to expand, traditional fact-checking methods struggle to keep pace. This growing gap raises concerns about the effectiveness of existing solutions in addressing misinformation at scale. While the previous manual method of fact checking remains accurate, it seriously lacks the ability to scale with the amount of information shared everyday as well as the speed at which it is consumed. These limitations highlight the need for alternative fact-checking methods—potentially leveraging the same AI-driven technologies that contribute to the problem.

Chapter 4: AI and its ability to detect misinformation

It was not long ago when people considered AI to be a thing of science fiction movies. The only people really giving it the time of day were mathematicians and visionaries who saw potential in the field. Nowadays, AI is becoming more and more prominent throughout our world. The first real recognition artificial intelligence got from the public was back in 1997, when IBM's Deep Blue deep learning model was able to beat world chess champion, Gary

Kasparov, demonstrating the capabilities of AI's capability to strategically think and effectively execute decisions. Fast forward to 2010, where deep learning rapidly advanced due to an influx of large-scale data collection across various domains. Big data and deep learning evolution allowed for strides in fields like natural language processing, autonomous vehicles and image recognition. Now in the 2020's deep learning based AI systems are actively reshaping the lives of people all over the world (Calmu.edu).

Today's AI systems are extremely fast, intelligent and scalable. AI algorithms can process information significantly faster than humans. Applications like ChatGPT have the ability to summarize entire news articles in seconds. Furthermore, artificial intelligence utilizes machine learning and deep learning techniques in order to identify complex patterns within data. Lastly, artificial intelligence has scaled very well over the past decade. As more data comes in, AI gets smarter. These systems have been able to analyze and learn from extremely vast datasets, which helps them to have a strong understanding of a wide range of use cases over several topics. Top AI companies are continuing to build up infrastructure to help scale these technologies as well as optimize the underlying algorithms to be more efficient. This emphasis on scalability gives us confidence that this technology will only get better over time.

To better understand how AI is making its decisions, it is important to understand the key points of the technology. Firstly, the largest part of generative AI is its natural language processing technology. This is essentially an algorithm that gives computers the ability to interpret and manipulate the human language. By training AI on conversations between humans, AI has been able to recognize patterns on how most appropriately to respond to certain word combinations (Aws.amazon.com). A simple way to understand this is comparing AI's learning of language to a young child. After repetition of hearing "Hello" from their parents and a response

of “Hello” back from other people, children pick up on this pattern because they have seen it multiple times. This is the exact same process used in order to understand language. It analyzes millions of conversations and is able to pick up on patterns and pick out which responses are most appropriate based on what it has learned from past conversations.

Machine learning is a subset of AI that allows computers to learn from data and improve their performance over time without being explicitly programmed. It works by training models on large amounts of information, enabling them to recognize patterns and make predictions. One of the most powerful techniques within machine learning is deep learning, which uses neural networks to process complex data structures.

Neural networks are inspired by the human brain, consisting of layers of interconnected nodes called neurons. These layers work together to identify patterns, making AI capable of tasks like speech recognition, image processing, and decision-making. When data is fed into a neural network, it passes through multiple layers, with each layer refining the information and improving the model’s accuracy. This is how AI systems, such as ChatGPT, can understand and generate human-like text or how self-driving cars can recognize traffic signs and obstacles.

With the vast amount of information circulating online, the rise of misinformation has become a significant challenge. AI-powered solutions, utilizing the technologies discussed above, have been developed to help detect and combat false or misleading content. These systems analyze text, images, and videos to identify inconsistencies, verify claims, and assess the credibility of sources. Because AI is excellent at recognizing patterns in large datasets, researchers have leveraged these techniques to combat one of today’s biggest challenges, misinformation.

One of the primary ways AI detects misinformation is through natural language processing, which enables algorithms to analyze the structure and meaning of written content. By training on vast datasets of factual and misleading information, AI models can recognize patterns in misinformation, such as exaggerated claims, emotionally charged language, or inconsistencies in reporting. Machine learning classifiers assign credibility scores to news articles, flagging potentially false information for further review.

Another technique involves fact-checking algorithms, which compare statements against verified databases of factual information. AI models can scan official sources, such as government websites, academic papers, and reputable news outlets, to determine the validity of a claim. This approach is used by automated fact-checking systems like ClaimBuster, which analyzes text in real time and identifies statements that require verification. The most used AI powered systems that try to tackle misinformation are:

- ClaimBuster: An AI-powered tool developed by the University of Texas at Arlington that automatically detects factual claims in political speeches, social media posts, and news articles. It flags statements that need verification, assisting journalists and fact-checkers in identifying misleading content.
- Google Fact Check Explorer: A tool that aggregates fact-checking results from reputable organizations like PolitiFact and Snopes, helping users verify claims quickly. AI helps match claims with existing fact-checks to streamline the verification process.
- Full Fact's AI Tools: The UK-based fact-checking organization Full Fact uses AI to monitor media and social platforms for misleading claims. Their AI tools detect

recurring misinformation, alerting fact-checkers and policymakers to emerging false narratives.

- Microsoft's Project Origin: A deep learning initiative aimed at verifying the authenticity of digital content, particularly images and videos. It uses cryptographic techniques and AI to trace content back to its original source, helping prevent the spread of manipulated media.
- Meta's AI-Based Fact-Checking: Facebook and Instagram employ AI-driven systems to detect misinformation by analyzing patterns of false content. These systems work alongside human fact-checkers to limit the reach of misleading posts and provide users with verified information.(Rand.org)

These AI solutions all face some challenges. There is a big limitation with bias in the fact-checking process. Bias can come from the humans that are part of the fact-checking workflow, and can even exist within the AI tools themselves. AI can be easily set up to bias certain points of views. AI models are also trained on data from our world, which has plenty of bias within it, and those same biases are projected onto the model and its outputs, which can lead to incorrect and even dangerous outputs from AI. Additionally, none of these solutions propose a good solution to topics that are very nuanced and or opinionated comments.

As misinformation tactics become more sophisticated, AI systems must continually improve to keep up. Future advancements will likely involve multimodal AI models, which analyze text, images, and videos simultaneously to detect misleading content more accurately. Additionally, integrating blockchain technology for content verification and crowdsourced fact-checking powered by AI can further strengthen the fight against misinformation. By leveraging machine learning, deep learning, and natural language processing, AI continues to

play a crucial role in preserving the integrity of information and ensuring that users have access to accurate and reliable content. While generative AI has created new challenges in terms of misinformation, it has also created the potential for solving this issue using the same technology. So how effective have these AI solutions been? As seen in the case studies, AI has shown promise in detecting misinformation but still faces limitations. In the next section, we will explore these successes and ongoing challenges in greater detail.

Chapter 5: Case Studies

Since the introduction of large language models and generative AI, companies and institutions have looked for ways to leverage this technology to help mitigate misinformation. One well-executed case study is Full Fact, the UK's independent fact-checking organization. Full Fact is the UK's independent fact-checking organization, has harnessed artificial intelligence to enhance the efficiency and reach of its misinformation detection efforts. By developing a suite of AI-powered tools, Full Fact aims to identify, verify, and challenge false information more effectively, thereby promoting accurate public discourse. (FullFact.org)

Established in 2009, Full Fact's mission is to combat the detrimental effects of misinformation on society. Recognizing that bad information can lead to harm, erode trust in democratic institutions, and skew public debates on critical issues, Full Fact has continually sought innovative solutions to address these challenges. In May 2019, in collaboration with Africa Check, Chequeado, and the Open Data Institute, Full Fact was awarded the Google.org AI Impact Challenge grant. This support enabled the organization to explore how machine learning could revolutionize fact-checking processes.

Full Fact has developed a comprehensive set of AI tools designed to streamline the workflow in the following ways:

1. Data Collection and Monitoring: The system aggregates data from various sources, including live television broadcasts, online news outlets, and social media platforms. Users can customize data collection parameters through an intuitive interface.
2. Claim Identification and Classification: Once data is collected, the AI parses the information into individual sentences to pinpoint checkable claims. These claims are categorized into types such as quantitative assertions (e.g., "GDP has risen by x%"), causal statements (e.g., "this policy leads to y"), and future predictions (e.g., "the economy will grow by z"). This classification aids fact-checkers in prioritizing which claims to investigate.
3. Claim Matching and Real-Time Alerts: The tools continuously compare new claims against a database of previously fact-checked statements. If a claim has been addressed before, the system alerts fact-checkers, enabling swift responses to recurring misinformation.
4. Live Transcription and Monitoring: During live events, such as political debates, the AI provides real-time transcriptions and highlights statements that may require verification. This feature allows fact-checkers to address inaccuracies promptly, ensuring that corrections are disseminated when they are most impactful. (FullFact.org)

Full Fact's AI tools have been instrumental in supporting fact-checking organizations worldwide. As of 2024, over 30 organizations across 16 countries utilize these tools to monitor public discourse and combat misinformation. The technology supports multiple languages, including English, French, and Arabic, with plans to expand further.

The effectiveness of Full Fact's AI-driven approach has been demonstrated during significant events. For instance, during the 2024 UK general election, the tools enabled rapid analysis of party manifestos and real-time fact-checking of live debates. Full Fact's solution significantly reduced turn around time of fact-checking, which helps to stop false claims at their roots. This capability ensured that voters had access to accurate information, thereby fostering informed decision-making.

While AI has significantly enhanced fact-checking capabilities, Full Fact acknowledges the limitations and challenges inherent in relying solely on automated systems. AI models can sometimes generate inaccuracies or "hallucinations," underscoring the necessity of human oversight to ensure contextual understanding and accuracy. Full Fact emphasizes that AI is a tool to augment, not replace, the nuanced judgment of experienced fact-checkers.

Looking ahead, Full Fact is committed to refining its AI tools and expanding their accessibility. The organization aims to foster a global collaborative network, enabling media outlets, civil society groups, and policymakers to better understand and address the misinformation landscape. Full Fact accomplishes this through their services as well as training courses they provide to help people get better at finding bad information. By continually enhancing their technology and sharing their expertise, Full Fact strives to create an information environment where accurate and reliable information prevails.

In summary, Full Fact's integration of AI into its fact-checking processes represents a significant advancement in combating misinformation. Through innovative tools and global collaboration, Full Fact is at the forefront of efforts to ensure that public discourse is informed by accurate and trustworthy information. While Full Fact has been able to successfully find a way to integrate AI as a tool for fact checking, the company has expressed to users to proceed with

caution when allowing AI to make decisions that are not validated by human experts. Their hesitation in allowing for a fully autonomous fact checking system implies that they do not believe this technology to be in a place where it can curate media all by itself. Building upon this idea, it's important to recognize that some organizations have already begun experimenting with fully autonomous, AI-driven fact-checking systems. One particularly compelling example emerges from the healthcare industry, where accurate information is crucial. In this field, AI systems are being developed and tested to independently verify medical facts and claims, aiming to improve the reliability of health information and ultimately support better decision-making for both professionals and patients.

AI's limitations in detecting health misinformation became evident in a June 2024 report by Derek Beres, which explored how chatbots like ChatGPT and Microsoft's Copilot often fail to provide accurate health-related advice. In addition to this analysis, health organizations and policymakers have responded by implementing stricter guidelines for AI-generated content, particularly concerning health advice. For instance, the World Health Organization has initiated collaborations with major tech companies to vet AI-driven health information. By integrating these broader institutional responses, we gain a more complete view of the efforts to mitigate AI shortcomings. Health-related advice is such a difficult domain because it is a field where even the most prestigious professionals do not always agree, so AI cannot be expected to always give great health-related information.

The problem extended beyond professional certifications. When asked about the supposed health benefits of coffee enemas, Copilot not only failed to clarify the lack of scientific evidence but also promoted links to purchase enema kits. Similarly, inquiries about turmeric supplements led to contradictory responses. To strengthen this analysis, it is important to discuss

countermeasures: organizations like Full Fact, HealthFeedback.org, and FactCheck.org have developed partnerships with platforms to monitor and correct health misinformation in real-time.

The proliferation of anti-vaccine rhetoric, fueled by AI-generated content and social media algorithms, has also contributed to declining vaccination rates. Since 2019, kindergarten immunization rates have dropped by 2%, and more than 8,000 schools in the United States are now at risk for measles outbreaks (CDC, 2024). These statistics reinforce the causal link between misinformation and public health risks.

Further complicating the issue, AI-generated content occasionally undermined its own advice. For example, when asked about credible alternative therapies for cancer, Copilot correctly stated that such treatments cannot cure cancer but then directed users to the Cancer Center for Healing in Irvine, California. This center promotes hyperbaric oxygen and ozone therapies, neither of which is FDA-approved for cancer treatment. Such inconsistencies illustrate how AI, despite its capacity to synthesize information, often fails to distinguish authoritative medical advice from fringe claims.

The risks associated with AI-generated health misinformation were further highlighted by an experiment conducted by Bradley Menz, a clinical pharmacist and AI researcher at Flinders University. Menz used an AI model to generate over 17,000 words of health-related blog posts containing vaccine and vaping misinformation. These posts were accompanied by 20 realistic, AI-generated images created in under two minutes. The content not only appeared credible but also included fabricated academic citations and testimonials, demonstrating how easily AI can be weaponized to spread false health narratives. Menz warned that such realistic misinformation could have "devastating" consequences, especially given the increasing reliance on online sources for health information (News Desk 2023).

The root of the problem lies in how AI models are trained. Large language models are built on vast datasets sourced from the internet, encompassing both credible research and unverified claims. Without the ability to critically assess the reliability of each source, AI models can inadvertently amplify pseudoscientific narratives. As Stanford Internet Observatory scholar Renée DiResta points out, there is a growing perception that AI-generated information is more trustworthy than social media content. However, users are often left to vet the accuracy of the information themselves, placing an undue burden on individuals to navigate complex health claims without expert guidance.

Efforts to address these shortcomings are ongoing but far from complete. While regulatory initiatives, such as President Biden's 2023 executive order mandating the watermarking of AI-generated content, represent progress, they do not address the fundamental challenge of ensuring accuracy in AI-driven health communications. Some researchers, like Wenbo Li from Stony Brook University, advocate for digital literacy programs aimed at teaching users how to critically evaluate chatbot-generated content. Li's work specifically focuses on training underserved communities, such as Black and Hispanic populations, to navigate generative AI technologies more effectively and identify credible health information.

Ultimately, while AI has proven to be a powerful tool for information dissemination, its current limitations underscore the importance of human oversight, particularly in sensitive domains like health care. The failures highlighted by Beres and Menz demonstrate that, without rigorous safeguards and continuous refinement, AI risks becoming not just a solution to misinformation but a contributor to its spread. This stands in stark contrast to the cautious, human-in-the-loop approach taken by organizations like Full Fact, reinforcing the notion that AI, while transformative, remains an imperfect guardian of truth.

Chapter 6: Ethical Implications of AI Misinformation Detection

Artificial Intelligence and machine learning have transformed many different avenues of life over the past decade. From self driving cars to helping draft up emails, there are a ton of use cases for this technology. The ability of AI to enhance these day to day tasks comes from its ability to recreate the way the world operates based on millions of data points. AI's pattern recognition algorithms are what helps it to accurately recreate the data points it has learned from. While these algorithms have proven to perform extremely well at what they were designed to do, there can be many unforeseen implications of this technology. AI's algorithms learn everything they can from the world that is already in place, so if that world has deep bias and unfairness, then the algorithms will pick up on those patterns and further push them onto society. One prime example of this is redlining – the process of refusing to grant people financial services because of living in a poverty labeled area. In 1933, Roosevelt created the Home Owners Loan Corporation, which was an institution set up to try and revive the failing housing market after the great depression. The group essentially looked at hundreds of cities and neighborhoods across the country and cross-referenced that data with risk of defaulting on loans. So they then used a mathematical algorithm to label each area as safe or risky for handing out loans. This initiative actually did improve the housing market but at a serious unethical cost. Many of the families in the “risky” zones were minorities who have been forced into these communities because of segregation and racism in the country. This algorithm essentially was telling banks to not give out loans to minorities which further separated them financially from the rest of the population, further burrowing them under the poverty line (Perry 2018).

This example from the past teaches the lesson that mathematical based algorithms will only recreate biases instilled in the data it is learning from rather than push the bar away from our previous mistakes. If AI is being used to detect misinformation, it has to develop some form of truth from the data given. So it will likely latch onto the most published information, rather than deciphering if it is coming from biased sources. Additionally, misinformation exists all over the internet. If the AI algorithms are learning from the internet, it means that AI algorithms can be trained on these false claims and would further push these false claims onto society.

Recent society has heavily debated the issue of freedom of speech versus moderation. Many people believe that America should have as little moderation as possible and there should be freedom of expression. Other Americans believe that there is not enough moderation and in turn harmful content is constantly being pushed onto society. This argument becomes even more complex with the introduction of AI into the mix. Some people may see it as an unbiased third party fact checker while others may immediately envision scenes from sci-movies like terminator. Both camps do have valid reasons for feeling this way, but either way the tension between the two sides has only grown stronger over the years and will continue to. As different political parties take office, it seems that this line is moved more one way or the other, which would make it extremely difficult for AI to actually be useful if the question it is trying to answer is extremely nuanced.

Another serious issue that would have to be addressed is accountability for promoting misinformation. In today's world news companies are expected to take full accountability for their mistakes in reporting. However in tomorrow's world, if AI is one day made mainstream in fact checking or even producing mainstream news-like content, who is responsible if misinformation is spread? Some might argue it is the developers fault for building a model with

flaws that allow it to spread misinformation. This is a difficult argument because while yes, those developers did create the AI, they are also dealing with an extremely complex technology and asking them to be well versed on all of the news topics may be asking them to wear too many hats. Some people may argue it is the users responsibility to handle misinformation. This argument must first make the assumption that users are fully aware that the analysis or content they are consuming is AI generated. However, the world is so fast paced nowadays – it is unlikely that people will go digging deeply in order to validate claims. This means that they will be avenues for the spread of misinformation. At the end of the day, they are just consumers and will not have the time to research every story. Lastly, people will argue the platforms are responsible for the spread of misinformation. This is the most sound argument. While platforms get criticized for allowing misinformation on their platforms, they are not intending for it to be spread, but if they introduce a misinformation detecting AI into their platform, then the accountability is solely in the hands of that platform. Many companies will see the speed and accuracy of some use cases and be tempted to use this technology, but they should also be fully aware of the downfalls and be willing to own up to their lack of diligence if misinformation is indeed spread through their implemented technology.

Another implication that is tied to any use case of AI, is the infringement of personal data. AI companies will only improve their models with more data, so they will always be looking for more. The more use cases AI is found to have, the more willing companies will be to sell data to these companies. Pushing any AI technology will always have the maybe unintended support for more and more data collection. One of the biggest ChatGPT models is trained on about 570 gigabytes of data, which equates to about 300 billion words (Menear 2024). When a training set is that big, it becomes almost impossible to audit exactly what data exists inside of

there, and that training set will only grow. Selling data has become one of the most lucrative ways, emphasized by Google's switch to becoming a "data company".

Another possible shortcoming of this technology would be the negative impact its failed results could have on society. To label something as true gives it immense power. An example of this is a conspiracy theory called *Pizza gate*. It was a fake story that a paedophilia ring involving the highest levels of the democratic party were operating out of a Washington pizza restaurant. These lies were believed by thousands of people and even convinced one man to drive hundred of miles and open fire after claiming he was trying to investigate the claims (BBC 2016). This story highlights the power of misinformation and how important it is to not spread lies that will cause society to erupt in harmful ways. If this technology was fully integrated and was even able to gain trust for great results over the years, imagine the damage even one false story could cause. Additionally, this technology may even increase the complexity of misinformation. People can find ways to slip their misinformation into the training data of these AI algorithms which may hinder its ability to detect misinformation.

Chapter 7: Challenges in Implementing AI Solutions

When implementing AI solutions for detecting misinformation, there are both practical and technical barriers that need to be overcome in order for this technology to become a core part of fact-checking. The challenges range from technical issues like context understanding and false positives, and extend to cost and environmental issues associated with maintaining these technologies. These factors significantly hinder stakeholder adoption across various domains, particularly in misinformation detection.

Contextual understanding means that AI systems can interpret information similarly to how humans do. Instead of viewing each piece of information separately, these systems make decisions by considering the bigger picture or overall context. A major challenge with Large Language Models (LLMs), such as ChatGPT, highlighted by researchers, is contextual understanding. Although developers aim to enhance contextual capabilities in AI, the technology currently remains insufficiently advanced. It is notably challenging to teach systems to grasp the subtleties of human language, emotional nuances, and implicit meanings that humans inherently recognize. AI often fails to perceive patterns in nuanced situations or fully understand the underlying reasons behind certain statements or events. Effective misinformation detection heavily relies on accurately interpreting context, yet AI models typically provide definitive answers rather than expressing uncertainty. This limitation highlights a significant obstacle to AI's effectiveness in combating misinformation. The only viable approach to enhancing contextual understanding involves extensive model training and continuous usage—both expensive processes that raise considerable financial and environmental concerns.

The financial implications associated with AI implementation are significant, involving software licensing fees, hardware upgrades, data acquisition, and consulting costs. These expenses can be prohibitive for smaller organizations, limiting AI's widespread adoption. This financial barrier is particularly concerning for misinformation detection technology, which necessitates diverse and inclusive participation to ensure unbiased and fair model development. Without broad-based engagement across socioeconomic strata, AI models risk reflecting the biases of affluent institutions exclusively, undermining their objectivity and efficacy.

In addition to initial implementation costs, ongoing operational and maintenance expenses present further financial challenges. Continuous infrastructure upkeep, software

subscriptions, and cloud computing expenses can rapidly escalate, making AI technology financially unviable for many organizations. Moreover, the uncertain return on investment exacerbates concerns. Major AI firms, including those behind ChatGPT, have publicly acknowledged their lack of profitability, causing investors to remain cautious regarding the sustainability of AI investments and their future economic implications (Quiroz-Gutierrez 2025).

Beyond technical and financial considerations, AI implementation faces substantial Information Technology (IT) challenges. Integrating AI technology into existing systems demands significant modifications to current IT infrastructures. Many organizations lack the robust technological framework necessary for seamless integration, requiring substantial upgrades or complete overhauls. It will take a lot of companies and institutions time to build up IT teams that are capable of implementing AI into their workflows. With the uncertain accuracy of AI in some domains, the question of whether or not to invest in the IT expansions necessary becomes very difficult to answer. Additionally, data privacy and security become paramount concerns, especially when handling sensitive or personal data. Organizations must adhere to stringent data protection regulations, complicating deployment and operational processes further. If companies want to implement AI solutions, they have to be able to navigate how AI's data infringements. It is very likely that a lot of consumers become skeptical of AI products even if they promise not to collect user data.

Environmental challenges represent another critical obstacle. Training and running AI models, particularly LLMs, consume substantial amounts of electricity, often generated from non-renewable sources such as coal, natural gas, and petroleum-based fuels. For instance, ChatGPT alone has been estimated to consume over half a million kilowatt-hours of electricity daily. This equates to approximately the daily energy consumption of 180,000 average U.S.

households, emphasizing the massive scale of energy demand posed by AI technologies (Strubell 2019). This extensive electricity usage contributes significantly to greenhouse gas emissions, thereby accelerating climate change and environmental degradation. Furthermore, AI models' high computational intensity requires the use of extensive data centers, typically cooled through energy-intensive methods, further exacerbating their environmental footprint. Addressing these environmental challenges requires companies to invest in renewable energy sources, like solar or wind power, or advanced, energy-efficient cooling systems and infrastructures. However, transitioning to greener technologies typically necessitates significant upfront investments, presenting additional financial barriers. Organizations must carefully balance technological ambitions with sustainability commitments, incorporating eco-friendly policies and practices. This involves not only infrastructural changes but also considerable investment in research and development to enhance energy efficiency and minimize ecological impacts. The integration of sustainability considerations increases both financial burdens and complexity, presenting another layer of difficulty for organizations attempting to leverage AI for misinformation detection.

Resistance from governments presents one of the primary barriers to adopting AI solutions, particularly those aimed at misinformation detection. Many governmental entities grapple with regulatory uncertainties, as existing frameworks often lag behind rapid technological advancements. Regulators face challenges determining how to classify AI-driven tools, establish accountability, and enforce compliance without stifling innovation. Governments are also wary of potential biases within AI algorithms, fearing that automated systems might inadvertently perpetuate or even exacerbate societal inequalities. Concerns about transparency and the explainability of AI decision-making further complicate regulatory acceptance, as

officials seek assurances that these systems will function equitably and transparently (AI Now Report 2018).

Platforms, including social media companies and tech firms, demonstrate resistance stemming from fears regarding liability and accountability. These concerns are particularly acute in sensitive areas like misinformation detection, where incorrect or biased AI decisions could have profound real-world consequences, impacting individuals' reputations, public trust, or even democratic processes. Platforms worry that integrating AI solutions could expose them to litigation or reputational damage if algorithms produce erroneous results. Additionally, they face operational challenges in maintaining effective human oversight and control over complex automated systems, raising further doubts about the feasibility of deploying AI solutions at scale without significant legal or reputational risks (Vaele & Binns 2017).

A prominent example of platform resistance to AI-driven misinformation detection is Facebook's experience during the 2020 U.S. elections. Despite deploying AI-based fact-checking systems and partnering with third-party organizations like PolitiFact and FactCheck.org, Facebook struggled to contain the rapid spread of false claims, particularly around voter fraud. Internal reports revealed that misinformation often bypassed automated filters due to algorithmic prioritization of engagement, and fact-check labels were inconsistently applied or delayed. Moreover, political pressure and public scrutiny limited Facebook's ability to aggressively moderate misleading content. This case illustrates how operational, reputational, and political pressures can undermine the effectiveness of AI-based fact-checking initiatives, highlighting the complexities platforms face in deploying such solutions.

Labor unions, civil society groups, and the general public often oppose the deployment of AI technologies due to concerns over job displacement and broader ethical implications.

AI-driven automation threatens jobs in various sectors, creating economic anxiety among workers who fear obsolescence. Civil society groups highlight ethical concerns, emphasizing issues such as privacy infringement, surveillance, and algorithmic fairness. Public mistrust is fueled by perceived opacity in how AI systems make decisions, leading many to question the fairness and impartiality of algorithmic judgments. This skepticism is heightened by high-profile cases where AI systems have shown biased outcomes or opaque decision-making processes, reinforcing apprehensions about the unchecked proliferation of AI technologies.

Collectively, these factors—governmental hesitation, platform fears, and societal opposition—create significant hurdles for the widespread adoption and successful deployment of AI systems for misinformation detection. The interplay of regulatory ambiguity, accountability fears, and ethical concerns underscores the complexity of navigating stakeholder interests in AI integration, ultimately slowing progress despite AI's considerable potential in combating misinformation.

Chapter 8: Performance Analysis

Building upon the theoretical foundations and case studies presented in earlier chapters, this section presents an experimental analysis of a large language model's (LLM) ability to classify the truthfulness of online articles. While existing tools like ClaimBuster and Full Fact utilize AI for misinformation detection, this chapter seeks to evaluate the standalone performance of an LLM using a custom scoring system against a labeled real-world dataset.

Methodology

To assess the model's classification ability, a prompt-based evaluation strategy was used. Articles were to a language model, and each article was assigned a credibility score between -1 and 1, where higher scores indicates a greater factuality score. This score was generated using a composite measure of four key criteria:

1. Factuality: Does the content align with verifiable data?
2. Logic: Are arguments coherent and free of contradictions?
3. Sentiment: Does the tone appear neutral or manipulative?
4. Bias: Does the piece appear ideologically skewed or present a balanced view?

These four criteria come from the research conducted from an AI startup company working on this issue. Each paragraph of every article was individually scored on these dimensions, and a weighted average was used to calculate a final article-level truthfulness score.

To evaluate the effectiveness of these classifications, the generated scores were tested against a labeled dataset containing 100 real-world news articles classified as either *True* or *False* based on established fact-checking sources. A decision threshold was calculated from -0.3 to 1.0 and model predictions were compared with the ground-truth labels to compute classification accuracy. It is important to mention that 100 is a generally small training model, but used purely because of financial constraints of the AI startup and the high cost associated with large language models. The size of the data is an undeniable weakness of these tests, and should and will be reconducted with much larger scale training and testing sets.

Results and Interpretation

The performance of the classifier at various thresholds is shown in **Figure 2.**, which plots Accuracy vs. Threshold.

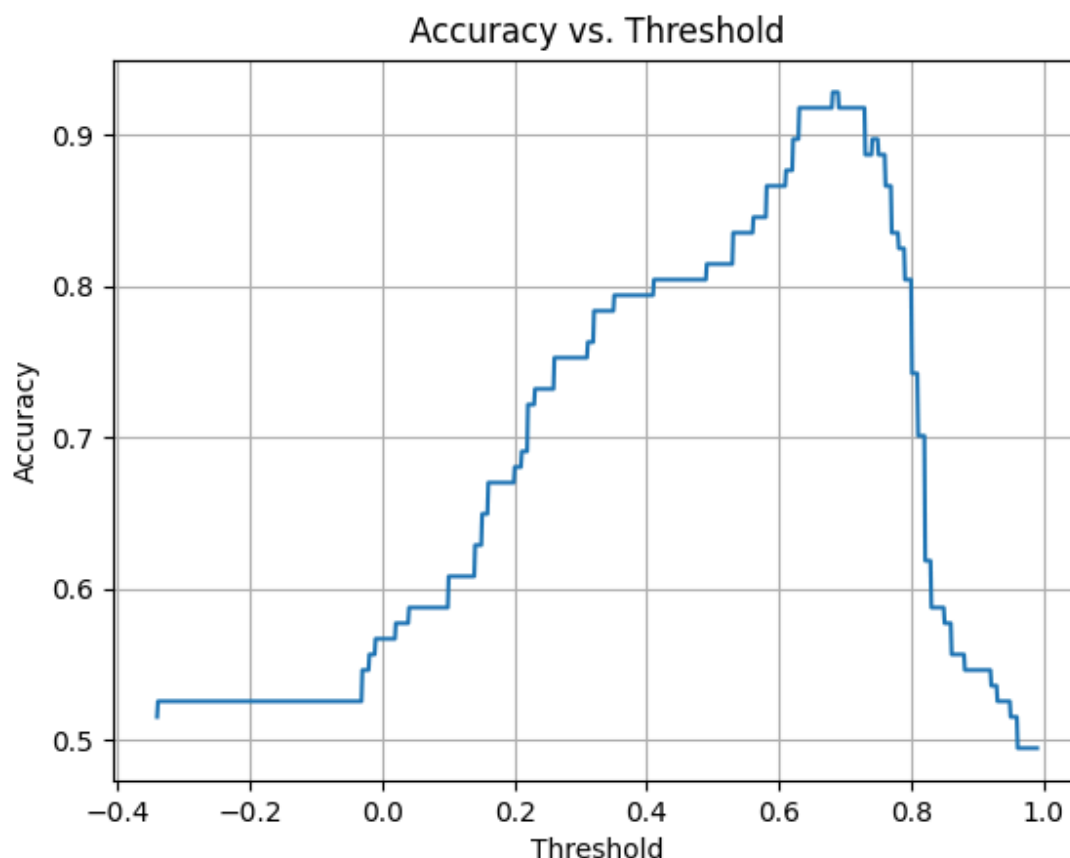


Figure 2. Line plot illustrating Accuracy vs Decision Threshold for LLM based classification model. This plot shows the change in accuracy at different selected thresholds for the classification model.

For clarification, the x-axis represents the threshold at which the model counts an article as truthful or not. For example, when the model claims that all articles with a truth score of 0.2 or higher is true, then the model is about 68% accurate, and when the model predicts any model with a truth score of 0.8 or higher is true, then it is about 75% accurate. There are also other ways of visualizing this data that are not present in this thesis. A confusion matrix at the highest

threshold would be interesting to look at in order to analyze where the model is making right and wrong choices by looking at the true positive, true negative, false positive and false negative. This would help to further analyze the model's true performance and paint a clearer picture. This insight would help to ensure the models accuracy across high-risk domains where false positives may be very costly. Understanding where the model is coming short helps to better tune it to become more robust.

As the graph illustrates, accuracy improves significantly as the threshold increases from 0.0 and peaks around 0.68–0.70, where classification accuracy exceeds 92%. This suggests that the model is most effective at distinguishing between true and false articles when a relatively conservative threshold is used—i.e., only confidently high-scoring articles are classified as true.

Beyond this peak, accuracy declines rapidly as the threshold becomes too strict. At very high thresholds (e.g., >0.9), the model becomes overly skeptical and misclassifies even truthful articles as false, while at very low thresholds (<0.2), it fails to detect false information effectively. This pattern reveals a sensitivity to threshold calibration, a common challenge in binary classification tasks with probabilistic models.

Interestingly, the model maintains near-random accuracy ($\sim 50\%$) at thresholds below 0.0, which aligns with expectations and confirms the classifier's reliability in distinguishing signal from noise.

To conclude, the results shown in figure 2, show the best possible results of a classification model that is trying to predict whether an article is true or false. These results mean that this model can be accurate when hyper tuning the decision threshold of the model, which helps to boost accuracy, but inhibits the robustness of the model as it is unclear whether more test sets will have the same domain.

Implications

The results from this test validate that LLM’s, when guided by structured prompt engineering and evaluated with proper scoring systems, can perform competitively in the domain of misinformation detection. With over 90% classification accuracy near optimal thresholds, this method holds promise as a scalable alternative to human fact-checking in low-risk environments or as a triage system to assist human reviewers.

However, this performance hinges on careful threshold calibration. A poor threshold can either lead to unchecked false positives or a suppression of truthful content. Moreover, the classifier’s scoring system assumes a relatively neutral domain; its effectiveness may vary across political or emotionally charged topics, as discussed in the implications chapter. So while this technology can be accurate in a vacuum, there is no guarantee that this technology can generalize well. The problem with this is that an organization could headline a 90% accurate model, fostering high trust amongst society, but if that model is then deployed into the real world and does not perform well, it would actually increase the credibility of misinformation if the model classified it as fact. This means that more stress testing needs to be done on this model in order to guarantee generalizable accuracy across a very wide range of domains.

Limitations

Despite the promising results, this experiment does not resolve the problem of ground truth ambiguity in nuanced or opinion-driven articles. Furthermore, because the LLM is trained on internet-scale data, its internal knowledge may itself be partially shaped by misinformation. Future work could explore integrating external fact-checking databases or consensus-driven

scoring systems to mitigate this issue. As mentioned in earlier chapters, organizations like politifact have presented frameworks where they utilize AI, but always have a human in the loop for real-time validation or feedback loops.

Additionally, the evaluation is dependent on the quality and diversity of the labeled dataset. Expanding to multilingual corpora and cross-topic articles (e.g., health, politics, science) would offer a more comprehensive understanding of model robustness. Furthermore, the presented test validity would be benefited heavily from additional accuracy metrics, like precision, recall, confidence intervals.

Conclusion

The performance analysis demonstrates that large language models, when structured and tested appropriately, can achieve high accuracy in classifying misinformation. This supports the hypothesis that AI can be a powerful tool in the battle against false information. Nonetheless, threshold selection, human oversight, and context awareness remain vital components for deploying such systems in real-world applications. While the empirical data from this study shows the possibility of this technology, humans guiding this tools remains non-negotiable to ensure total accuracy of results, as even a 90% accurate model, which is good in terms of a classification model, will leave 10% of misinformation to exist which can have serious consequences as pointed out in subsequent chapters.

This chapter underscores the growing potential of AI for scalable, semi-automated fact-checking, while also recognizing the technical and ethical hurdles that must be navigated to ensure these tools do more good than harm.

Chapter 9: Future Directions and Recommendations

Development of Multimodal Misinformation Detection Systems

One key avenue for future research is the development of multimodal AI systems that integrate text, image, audio, and video analysis to detect misinformation. Multi-modal approaches are helpful in mitigating biases because it allows for multiple models to make predictions which can help mitigate some bias issues mentioned earlier in the paper. While most current systems focus primarily on natural language processing (NLP) for text analysis, misinformation in the digital age increasingly relies on diverse media formats such as deepfake videos, altered images, and synthetic audio.

Recent studies have shown that multimodal systems provide significantly higher accuracy in detecting misinformation because they can cross-validate information across different content types (Khan et al., 2023). Companies like Microsoft and Meta have already begun experimenting with these systems for content moderation but face challenges with computational costs and scalability (Gupta et al., 2022).

The limitations in the current infrastructure in AI are a current roadblock, but this will not always be the case. Companies like Google have begun investing in new forms of energy like nuclear to back the computational costs of this technology. In addition, companies like DeepSeek have found ways to improve algorithmic efficiency of generative AI (Bain.com). This push to make the models more efficient is also another front to help get over the computational cost and stability roadblock that multimodal AI systems face.

Incorporation of Explainable AI (XAI) Frameworks

A major limitation of current AI misinformation detection tools is their "black-box" nature, where decisions are made without user-facing explanations. Explainable AI (XAI) frameworks provide transparency, allowing users to understand why a particular piece of content has been flagged or verified.

Transparency has been shown to foster user trust and facilitate more effective user engagement with AI tools (Doshi-Velez & Kim, 2017). Particularly in politically charged environments, the ability to present clear rationales for decisions will be crucial for ethical AI deployment.

Companies like OpenAI have been experimenting with new models called “reasoning models”. These new models explain to the user its “rationale” for making decisions. Below is an example of chat with OpenAI’s most advanced reasoning model, o3.

User prompt: The world is flat.

o3 output: The user says, "The world is flat." I'm interpreting that either as a statement of belief or possibly as a test or invitation for a debate. I'll respond by politely correcting the claim, explaining Earth's spherical shape, and providing evidence for that. I'll maintain a friendly and informative tone, considering their interest in more grounded topics like financial planning, while avoiding getting overly technical. No need to search for additional data—we know the Earth isn't flat from existing knowledge.

These reasoning models perform surprisingly well in terms of accuracy, but they also foster huge prospects in terms of selling AI to users. The ability to explain the reasoning behind an argument adds validity to its claims. Detecting misinformation using AI is a great use case for

this technology. The example conversation above presented a false claim to the AI as fact. The AI system decided that this was a false claim and decided the best thing to do was correct the user. Even more interestingly, it was able to reason whether or not it needed to receive additional data. The AI system decided that it was so confident the earth was not flat, that there was no need to expend more resources and try to collect more data. Below is a prompt that challenges AI with a more ambiguous prompt.

User prompt: *Aliens are real.*

o3 output: *The user says "Aliens are real," which seems like a statement rather than a question. They might want to discuss their belief or explore evidence. I could respond by asking if they have specific evidence or sources they're referring to, while offering the scientific perspective. There's always news around extraterrestrial life, like recent congressional hearings or NASA's UAP group work. I could check for the latest updates on the topic so I can provide a current response.*

To respond thoughtfully... (Full output can be found in the appendix.)

The model then goes on to inform the user that there is no current evidence to support the claim that aliens are real. This example chat shows an example where the AI system feels it needs to get up to date on this information, so it searches the web and regurgitates up to date information back to the user. These reasoning models not only provide accurate feedback to the user but also showcase the ability to learn with the most up to date information by surfing the web.

This technology helps address the limitations of user trust around AI. As explained in subsequent chapters, LLM are black box models, meaning their reasonings for decisions are

abstracted from user's. This technology introduces the ability to use AI for answers and understand its decision making process for these answers. This technology is still a work in progress, but it still shows the potential for addressing the limitations of user trust when it comes to decisions made by AI. One important thing to note for this technology is the fact that it becomes more dangerous the more validity it has. The above use cases showed cases where the AI is correct in its explanation so it works well, but there's a possibility that the AI uses misinformation to reason through its output. This is a serious potential downside to these reasoning models and will be a big hurdle that these models will need to get over.

Ethical AI Design and Bias Mitigation

As discussed in subsequent chapters, AI models are only as unbiased as the data they are trained on. This presents challenges, especially for global misinformation detection systems operating across cultures and languages. A lot of the data that has been collected from the world comes from places with deep rooted biases. Groups that have been oppressed and pushed down by society are underrepresented in data, which of no fault of their own, can lead them to be underrepresented in models trained on that data.

This underrepresentation creates significant practical and ethical concerns. When AI models used for misinformation detection rely on biased data, the consequences can disproportionately affect already marginalized groups. For instance, automated systems intended to flag harmful or false content can unintentionally amplify discrimination if they fail to accurately recognize culturally specific language or context. A study by (Noble 2018) demonstrated how algorithmic search systems, particularly Google, often returned highly stereotypical or offensive results when users searched for terms related to Black girls and

women. For example, a search for “Black girls” frequently yielded hypersexualized or pornographic content, revealing how commercial search algorithms prioritize click-driven relevance over fairness or social accuracy. This finding is particularly alarming because it shows that algorithmic systems trained on unmoderated data can not only reflect but also amplify societal prejudices. In the context of misinformation detection, this suggests that AI tools, if left unchecked, could inadvertently validate and spread harmful stereotypes under the guise of objectivity, reinforcing structural inequalities rather than mitigating them.

Moreover, if AI-driven misinformation detection systems reflect biases, users may perceive these systems as inherently unjust, eroding public trust in AI technology and potentially exacerbating digital divides. Such mistrust could lead affected communities to reject valid information flagged by AI systems, ultimately undermining the effectiveness of AI-based solutions against misinformation.

Future AI development should incorporate fairness-aware machine learning algorithms that specifically address bias in training data (Mehrabi et al., 2021). Additionally, culturally adaptive models that are sensitive to regional norms and languages must be prioritized.

To achieve these goals, machine learning engineers and researchers must adopt rigorous methodologies for detecting, understanding, and mitigating bias at all stages of AI development. Techniques such as counterfactual fairness, which assesses how AI decisions would differ under alternative demographic scenarios, can provide deeper insights into hidden biases. Furthermore, the employment of tools like adversarial de-biasing—where AI models are explicitly trained to reduce biases identified by a competing adversarial algorithm—has shown promise in creating fairer machine learning outcomes (Zhang et al., 2018).

In addition to algorithmic adjustments, the datasets themselves must be expanded to more inclusively represent global populations. Initiatives aimed at curating diverse and culturally rich datasets can help mitigate representation gaps, ensuring that marginalized or minority voices are adequately captured. This includes efforts such as involving diverse communities directly in data collection and validation processes, an approach known as participatory AI. This approach enables underrepresented groups to actively contribute to training data, providing a pathway for greater accuracy, fairness, and authenticity in the models produced (Sloane et al., 2020).

Ethical design must also emphasize transparency and accountability. Clear explanations of how AI systems detect misinformation, along with open reporting on model performance across different demographic groups, help build public trust. Regular external audits of these systems by independent third parties should become standard practice, ensuring continuous improvement and accountability for bias mitigation efforts.

Finally, interdisciplinary collaboration is essential. Ethicists, sociologists, cultural anthropologists, and community representatives must be included in the development teams alongside technical specialists. By incorporating perspectives beyond traditional engineering roles, developers can more effectively identify potential ethical pitfalls and proactively design solutions that are culturally aware and socially responsible.

AI Literacy and Public Education Programs

While technological solutions are critical, misinformation is fundamentally a human problem exacerbated by a lack of media literacy. As highlighted by Wenbo Li's research (Li 2024), underserved communities are disproportionately affected by misinformation due to lower levels of digital literacy. As AI becomes more prevalent and accessible to people around the

world, education of AI as well as education in general will need to make some adjustments in order to prevent AI misuse and to help understand the limitations of AI.

Current educational paradigms often do not adequately address AI literacy, leaving many individuals unequipped to critically evaluate AI-generated information. As AI-generated content becomes indistinguishable from human-generated content, the need for robust educational frameworks focused on AI literacy becomes increasingly urgent. Educational programs must incorporate foundational knowledge about how AI works, its potential biases, its strengths, and its limitations. Studies suggest that people who understand basic AI concepts are less likely to be misled by misinformation or overly reliant on automated systems for critical decisions (Long & Magerko, 2020). Therefore, governments, educational institutions, and technology companies should collaborate in developing comprehensive educational resources that specifically target AI literacy from an early age.

Children of new generations will be raised in a very different environment than any of the generations before them. They will essentially have the smartest tutor of all time in their pockets with them wherever they go in the form of AI. This presents both great opportunities and catastrophic consequences.

On one hand, AI can significantly democratize access to education by providing personalized learning experiences tailored to individual needs, abilities, and interests. Adaptive AI-powered platforms have already shown great promise in enhancing learning outcomes, particularly among students with diverse learning styles or special educational needs (Baker & Siemens, 2014). However, reliance on AI without proper critical thinking skills could lead students to trust the output of AI tools without questioning their accuracy or ethical implications.

The risk of developing an overreliance on AI-generated information makes it essential to emphasize critical thinking and skepticism as central components of future educational programs. Curricula must prioritize teaching students how to identify and scrutinize the authenticity of digital content, evaluate sources, and recognize signs of AI-generated misinformation. In practice, this means equipping students not only with the technical understanding of AI systems but also with the intellectual toolkit necessary to discern credible from deceptive information in real-world scenarios.

Furthermore, equity in AI education is vital. Historically underserved communities are particularly vulnerable to misinformation due to systemic disparities in access to quality education and digital resources. Educational programs must intentionally focus on bridging these gaps, providing AI literacy education and resources explicitly designed to support disadvantaged communities. Initiatives such as localized workshops, community-based digital literacy programs, and free online AI-literacy modules tailored to different cultural and linguistic contexts can help mitigate this digital divide (L, 2024).

Educators themselves must also be supported through training programs designed to enhance their own understanding of AI technologies. Teacher training initiatives should equip educators with both theoretical and practical knowledge of AI, empowering them to effectively guide students in navigating the complexities of AI-driven environments. Moreover, incorporating real-world applications of AI into classrooms can foster familiarity and reduce apprehension surrounding AI technologies among both educators and students.

Lastly, broader public education campaigns must accompany formal school-based education. Governments and public organizations should engage in continuous public outreach through social media, public service announcements, and community engagements, reinforcing

core AI literacy principles across diverse age groups and socio-economic segments. The ultimate goal is cultivating a public that is not only comfortable interacting with AI technology but also critically aware of its limitations and ethical challenges.

Educational institutions and policymakers should develop and implement comprehensive AI literacy curricula integrated at all educational levels, from primary school to adult education. These curricula should emphasize critical thinking, source verification, and practical AI knowledge. Special attention must be given to underserved communities through targeted initiatives and resource allocation to prevent exacerbating existing educational inequalities. Moreover, teacher training and public education campaigns should be prioritized to foster an informed, critical, and AI-literate society capable of effectively responding to the rapidly evolving information landscape.

Human-in-the-Loop Systems for High-Stakes Domains

As demonstrated in the case study on health misinformation, fully autonomous AI systems often fail in domains requiring nuanced judgment. Future misinformation detection systems should prioritize human-in-the-loop (HITL) frameworks, where AI assists but does not replace human experts.

High-stakes domains, such as healthcare, law, finance, and electoral processes, demand a level of judgment and contextual understanding that current AI models alone cannot consistently provide. Misinformation in these domains is also particularly dangerous because of how quickly it can manifest. These are all fields that take a lot of domain knowledge to understand, so the average person is not very well versed. This makes it much easier for people to be misled and in turn manifest this misinformation through their social circles. Human-in-the-loop (HITL)

systems are uniquely effective in this domain because it allows for someone with domain expertise to vet misinformation detection results. A misleading piece may be very obviously false to an expert but an average person with little to no domain knowledge may not know any better than to believe the piece. HITL systems integrate human oversight into AI processes to improve decision-making, accountability, and trustworthiness. Rather than replacing human judgment, HITL systems leverage AI's speed and scalability for initial filtering and detection, allowing human experts to focus on complex, ambiguous, or ethically sensitive cases.

In practice, a HITL system for misinformation detection could function by having AI tools rapidly scan and classify vast quantities of online content, identifying potentially problematic or uncertain cases. These cases are then escalated to trained human reviewers who apply nuanced analysis, cultural sensitivity, and ethical judgment to confirm, clarify, or correct the AI's initial assessment. This collaborative model allows AI to manage routine tasks effectively, while human experts handle the nuanced evaluation and decision-making processes that require critical reasoning and ethical sensitivity (Amershi et al., 2019).

Moreover, HITL systems enhance the transparency and accountability of misinformation detection processes. If an autonomous AI incorrectly flags valid content or fails to detect harmful misinformation, assigning responsibility can become unclear. In contrast, human-in-the-loop approaches ensure clear accountability, as final decisions involve human oversight. This clarity not only improves the credibility of misinformation detection outcomes but also reduces potential liability concerns for organizations implementing these systems.

Additionally, HITL frameworks facilitate continual improvement of AI models. Human decisions can serve as training data to refine and update AI algorithms, thereby progressively reducing errors and improving AI's ability to handle ambiguous situations. Regular feedback

from human reviewers also helps identify systemic biases or limitations within AI systems, guiding adjustments and iterative improvements (Holstein et al., 2019).

For HITL systems to be effective, human reviewers must be adequately supported. This includes providing extensive training in identifying misinformation, understanding cultural contexts, and recognizing algorithmic biases. Training programs should regularly update reviewers on emerging misinformation trends, technological advancements in generative AI, and evolving ethical considerations. Organizations implementing HITL systems must also ensure humane working conditions and manageable workloads, as cognitive overload could compromise the quality of human oversight and decision-making.

Despite the clear advantages, implementing HITL systems at scale faces challenges, such as operational costs, speed of human review relative to AI-only systems, and the risk of introducing human biases into AI-assisted processes. Therefore, careful operational planning and robust quality control measures are necessary. Organizations must balance AI's efficiency with the effectiveness and ethical oversight offered by human involvement, clearly defining procedures for when human reviewers must be involved and establishing metrics for performance evaluation.

Future misinformation detection initiatives in high-stakes environments should mandate human-in-the-loop integration, clearly defining human oversight roles. Organizations should invest in extensive training, regular review, and refinement cycles for both human reviewers and AI systems. HITL protocols must emphasize transparency, accountability, and bias mitigation to ensure ethical and accurate decision-making, ultimately fostering public trust in AI-driven misinformation detection systems.

Conclusion

The future of combating misinformation with AI represents both an extraordinary opportunity and a profound responsibility. As AI systems become increasingly sophisticated, they are capable of analyzing not just text but also images, audio, and video, so the potential for effective misinformation detection expands significantly. However, with this promise comes substantial ethical and practical challenges. Multimodal detection systems, while technologically powerful, must overcome substantial hurdles related to scalability, computational efficiency, and sustainability. The ongoing innovation in computational infrastructure and algorithmic efficiency offers hope, but these technological strides must always be aligned with ethical imperatives.

Explainable AI (XAI) frameworks further underline the importance of transparency in building public trust. AI's ability to provide clear, rational explanations for its decisions not only enhances accountability but also engages users in a meaningful dialogue about the content they consume. The emerging developments in reasoning models illustrate the potential of AI to communicate its thought processes, thereby reducing misunderstandings and reinforcing trust. Nonetheless, transparency alone does not guarantee trustworthiness; AI must also be designed to actively mitigate bias.

Bias mitigation and ethical AI design underscore that technology is never neutral, as it reflects and can amplify societal inequities. As AI tools grow in power and influence, the critical necessity of building fairness into their core becomes clear. Future advancements must integrate diverse perspectives and rigorous fairness methodologies at every stage of AI development. By prioritizing cultural inclusivity and participatory approaches, AI can not only detect misinformation more accurately but also do so equitably, reducing rather than exacerbating societal divides.

Yet, technology alone cannot fully resolve the misinformation crisis; human factors remain central. Effective misinformation detection must be supported by comprehensive AI literacy initiatives. Educating future generations about the strengths, limitations, and ethical implications of AI is essential. As society increasingly relies on AI-driven systems, ensuring citizens are equipped with critical thinking skills to navigate digital information becomes imperative. Without this foundation, even the most advanced AI technologies risk being misunderstood, misused, or mistrusted.

Finally, human-in-the-loop systems offer a compelling path forward for deploying AI responsibly in high-stakes domains. Rather than replacing human judgment, AI should enhance human decision-making, enabling rapid processing of vast data while preserving nuanced human oversight. Ensuring clear accountability, ethical scrutiny, and continuous improvement through human feedback provides a robust framework for sustainable, trustworthy AI solutions.

In closing, addressing misinformation through AI is not merely a technological endeavor, it is a deeply human one. Future strategies must holistically integrate technological innovation with ethical integrity, human judgment, and societal education. Achieving this balance represents the next great frontier in the pursuit of truth in our digitally interconnected world.

Chapter 10: Conclusion

The widespread proliferation of misinformation in the digital era represents one of the most urgent challenges facing contemporary societies. Traditional methods of detection and prevention like manual fact-checking, government and institutional policy interventions, and media literacy efforts, have proven insufficient in an age where information can be generated and disseminated instantaneously. This thesis set out to explore whether artificial intelligence (AI),

particularly systems leveraging natural language processing (NLP) and machine learning (ML), could offer scalable, accurate, and ethically sound solutions to combating misinformation.

Chapter 1 framed the core problem and established the central research question: Can artificial intelligence serve as a viable solution to the global misinformation crisis, and if so, what are the ethical, technical, and practical considerations that must be addressed to implement such systems responsibly? Subsequent chapters sought to address this question through a comprehensive examination of misinformation's evolution, current limitations in countermeasures, the technical promise of AI, and both its real-world applications and constraints.

Chapter 2 detailed the historical evolution of misinformation, tracing its roots from ancient propaganda campaigns to modern algorithmically amplified disinformation. This historical perspective highlighted that while the phenomenon is not new, the scale, speed, and scope of modern misinformation are unprecedented, especially given the rise of social media platforms and algorithmically curated content. The chapter underscored the psychological persistence of misinformation, even after correction, and introduced the necessity of preemptive detection strategies that can keep pace with the velocity of digital content.

Chapter 3 reviewed existing, predominantly manual solutions to misinformation, including independent fact-checking organizations, social media platform policies, and government-led educational initiatives. While accurate and well-intentioned, these methods are labor-intensive, slow, and prone to political pushback or enforcement inconsistencies. The rise of generative AI has made it possible to create realistic fake content at scale, rendering traditional reactive strategies insufficient. This served to justify the pivot toward automated, AI-driven approaches.

In Chapter 4, the thesis examined the core AI technologies underpinning misinformation detection. By leveraging NLP, deep learning, and supervised classification models, AI systems can be trained to identify linguistic and contextual markers of misinformation. Tools such as ClaimBuster and Google’s Fact Check Explorer already employ these methods to varying degrees of success. Importantly, the chapter explored how these models function, including their reliance on training data and their ability to process and score content based on factual consistency, logical coherence, emotional tone, and ideological bias.

Chapter 5 presented case studies that demonstrated the real-world implementation and limitations of AI fact-checking systems. The analysis of Full Fact illustrated how AI can meaningfully assist human fact-checkers by automating claim detection and alerting reviewers to previously verified claims. Conversely, the case of AI-generated health misinformation, particularly when AI tools suggested pseudoscientific treatments on equal footing with medically rigorous advice, exposed the risks of deploying AI systems in high-stakes, nuanced domains without sufficient safeguards. These contrasting cases reinforced the importance of maintaining human oversight and limiting the autonomy of AI in critical contexts.

Chapter 6 considered the ethical implications of AI-based misinformation detection, particularly concerns around bias, accountability, transparency, and data privacy. It emphasized that AI systems are only as objective as the data and design choices behind them. Without deliberate bias mitigation, AI can perpetuate and even amplify systemic inequalities, particularly when deployed across diverse cultural contexts. The chapter also raised critical questions about responsibility, should a system generate and propagate misinformation, who is held accountable: the developers, platforms, or end users? These questions remain unresolved but are vital in shaping future policy frameworks.

Chapter 7 examined the practical and technical challenges of AI implementation. Contextual understanding remains a key limitation—AI models struggle with ambiguous, ironic, or culturally specific language, and their accuracy diminishes significantly in edge cases or emotionally charged content. The chapter also discussed financial, infrastructural, and environmental costs associated with scaling AI, particularly large language models, raising concerns about equity, sustainability, and global accessibility. Furthermore, it explored institutional resistance to AI adoption, from government skepticism to public fears over job displacement and surveillance.

Chapter 8 offered empirical evidence of AI’s performance through a custom scoring system and threshold analysis using a labeled dataset. The model demonstrated strong predictive power, achieving over 90% classification accuracy under optimal conditions. However, its effectiveness was dependent on careful threshold calibration and a relatively neutral domain, confirming that AI’s utility is not universal but context-bound. The findings suggest that while AI can function as a scalable screening tool or triage mechanism, human review remains indispensable—particularly for edge cases, complex topics, or sensitive social domains.

Finally, Chapter 9 and the Future Directions sections synthesized strategic pathways for improving AI-based misinformation detection. These included the development of multimodal detection systems to address the diversity of misinformation formats, the integration of Explainable AI (XAI) frameworks to foster transparency and user trust, the incorporation of ethical and fairness-aware algorithms, and the design of AI literacy programs to ensure responsible use by the public. The discussion also strongly advocated for human-in-the-loop (HITL) systems, which strike a balance between speed and accuracy by combining AI’s processing capabilities with human contextual judgment.

Across these chapters, the thesis affirmed the core proposition: AI, while imperfect, holds potential as a scalable and effective tool for combating misinformation, provided it is deployed with transparency, ethical foresight, and ongoing human involvement. The findings suggest that AI should not be seen as a replacement for human discernment, but rather as a force multiplier that can augment and streamline fact-checking workflows. A main takeaway is that as these models become more accurate over time, their validity

The broader implications of this research are both urgent and far-reaching. As societies increasingly rely on digital platforms for news, education, and decision-making, the integrity of information becomes a cornerstone of democratic resilience, public health, and social cohesion. Misuse or over-reliance on opaque, unregulated AI systems risks deepening societal divisions and spreading harm, while thoughtful integration of AI into misinformation detection ecosystems offers a path toward greater informational clarity, equity, and trust.

In conclusion, the battle against misinformation is not solely a technological one, it is also ethical, political, and deeply human. AI systems must be built not only to detect falsehoods but to do so with transparency, fairness, and accountability. Only through the collective efforts of technologists, policymakers, educators, and civil society can the promise of AI be realized in service of truth. This thesis contributes to that vision by offering a foundational exploration of AI's current and potential role, its strengths, its shortcomings, and the very long and precise roadmap necessary to responsibly harness its power for the public good.

References

- Alkam, J., & DiGiacomo, F. (2024). *Advancing policy to foster K-12 media literacy*. National Association of State Boards of Education.
- Amazon Web Services. (n.d.). *What is natural language processing?*. Retrieved [date], from <https://aws.amazon.com/what-is/natural-language-processing/>
- Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2019). *Guidelines for human-AI interaction*. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). ACM.
- Bain & Company. (2025). *Efficiency improvements in AI infrastructure*. Retrieved from <https://www.bain.com>
- Baker, A., & Fairbank, V. (2022, Nov 14). *The Truth in Journalism Fact-Checking Guide*. Toronto Metropolitan University School of Journalism.
- Baker, R., & Siemens, G. (2014). *Educational data mining and learning analytics*. In J. Larusson & B. White (Eds.), *Learning analytics* (pp. 124-139). Springer.
- BBC News. (2016, Dec 2). *Pizzagate: The fake story that shows how conspiracy theories spread*.
- Brodsky, S. (2024, Aug 4). *Alex Jones damages crush Sandy Hook conspiracies*. *The Wrap*.
- Centers for Disease Control and Prevention. (2024). *Vaccination coverage and exemptions among kindergartners—United States, 2023–24 school year*. <https://www.cdc.gov/schoolvaxview/data>
- CISA. (2023). *Disinformation, misinformation and malinformation (DMM): An introduction*. Cybersecurity and Infrastructure Security Agency.
- Commons Library. (2024). *Misinformation, disinformation and malinformation: Definitions and examples*. House of Commons Library.
- Department of Homeland Security. (2021). *Media literacy and critical thinking online*. <https://www.dhs.gov/media-literacy>
- Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning* (No. arXiv:1702.08608). arXiv.
- FactCheck.org. (n.d.). *Our process*. Retrieved April 27, 2025, from <https://www.factcheck.org/our-process/>
- Flinders University News Desk. (2023, Nov 14). *Medical researchers find AI fails pub test*. <https://news.flinders.edu.au/blog/2023/11/14/medical-researchers-find-ai-fails-pub-test/>
- Full Fact. (n.d.). *About our AI tools*. Retrieved April 27, 2025, from <https://fullfact.org/ai/about/>
- Gupta, A., Singh, N., & Kumar, M. (2022). *Multimodal misinformation detection: Challenges and opportunities*. *Information Processing & Management*, 59(6), 103089.
- Holstein, K., Vaughan, J. W., Daumé III, H., Dudik, M., & Wallach, H. (2019). *Improving fairness in AI systems: What do developers need?* In *CHI '19 Extended Abstracts* (Paper 259). ACM.

- Khan, A., Singh, R., & Jain, S. (2023). *Multimodal deep learning for misinformation detection*. *IEEE Transactions on Knowledge and Data Engineering*.
- Konopliov, K. (2024, May 7). *The rising threat of deepfakes: Detection, challenges, and market growth*. *Liminal*. <https://liminal.co/articles/rising-threat-of-deepfakes/>
- Li, W., Flores, R., & Hernández, P. (2024). *Empowering underserved communities with AI literacy*. *Journal of Information Literacy*, 18(1), 123–140.
- Long, D., & Magerko, B. (2020). *What is AI literacy? Competencies and design considerations*. In *CHI '20 Extended Abstracts* (pp. 1–8). ACM.
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). *A survey on bias and fairness in machine learning*. *ACM Computing Surveys*, 54(6), 1–35. <https://doi.org/10.1145/3457607>
- Menear, H. (2024, Mar 6). “Big data” isn’t big enough to train generative AI. *Interface Media*.
- Meta. (2025, January). *More speech, fewer mistakes*. Retrieved from <https://about.fb.com/news>
- Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.
- Novet, J. (2025, Jan 10). *Zuckerberg says Biden administration pushed Meta ‘super hard’ to take down vaccine content*. *CNBC*. <https://www.cnn.com>
- Pangotra, V. (2024). *Social-media algorithms: The black-box driver of misinformation*. *CyberPeace Institute Blog*.
- Perry, A., Rothwell, J., & Harshbarger, D. (2018, March 20). *The devaluation of assets in Black neighborhoods: The case of residential property*. *Brookings Institution*. <https://www.brookings.edu/articles/devaluation-of-assets-in-black-neighborhoods/>
- Quiroz-Gutierrez, M. (2025, Jan 30). *Why 75% of businesses aren’t seeing ROI from AI yet*. *Fortune*.
- RAND Corporation. (n.d.). *Fighting misinformation with AI*. Retrieved [date], from <https://www.rand.org>
- Ross, V. (2011, July 1). *Lingering lies: The persistent influence of misinformation*. *Scientific American*. <https://www.scientificamerican.com/article/lingering-lies/>.
- Sifuentes, B. (2019, Mar 22). *The propaganda of Octavian and Mark Antony’s civil war*. *World History Encyclopedia*. <https://worldhistory.org/article/1474/>
- Sloane, M., Moss, E., & Sanchez, J. (2020). *Participatory approaches to machine-learning systems*. *Patterns*, 1(2), 100017.
- Socialinsider. (2023). *Social-media benchmarks 2025 report*. <https://www.socialinsider.io/social-media-benchmarks/>
- Strubell, E., Ganesh, A., & McCallum, A. (2019). *Energy and policy considerations for deep learning in NLP*. *ACL 2019*.
- Vaele, R. V., & Binns, R. (2017). *Fairness and accountability of automated decision-making: A legal perspective*. *Computer Law & Security Review*, 33(3), 308–321.

Weiner, M. (2024). *More than a century of antisemitism: Russia's disinformation playbook*. U.S. State Department Global Engagement Center.
state.gov

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 335–340). ACM.

Appendix

Continuation of ChatGPT output from page 39:

It's best to gather the most up-to-date information on extraterrestrial discoveries since new evidence always emerges. There have been discussions from recent years, like NASA's UAP panel and possible biosignatures from exoplanets. I'll aim to summarize findings for the user, offering scientific perspectives on the topic. Since they seem assertive, I'll clarify what they mean by "Aliens are real" and gently ask for their source while discussing the current state of the evidence.

Searched the web

- 2025 pentagon uap report january 2025
- NASA 2024 UAP independent study team final report September 2024
- Latest news & breaking headlines
- Disclosure Diaries
- NASA

The user's statement "Aliens are real" feels like a bold claim, but it's very open-ended. I'll respond by summarizing current scientific findings—NASA hasn't confirmed any extraterrestrial life, and while JWST hints at possible biosignatures, nothing is verified. UAP reports are intriguing, but they don't confirm alien origins either. I may also ask the user for any evidence they have while maintaining a respectful, open tone, since they seem confident in their claim. I'll craft the answer carefully and use sources for clarity.