A Survey on Misinformation Prevention and Detection methods in Large Language Models

Anonymous submission

Abstract

The rapid advancement of large language models (LLMs) has significantly impacted various fields within natural language processing (NLP). However, the issue of misinformation has become increasingly prominent, necessitating urgent solutions. Recent studies have categorized misinformation into two types: unintentional misinformation, often resulting from hallucinations, and intentional misinformation, which is deliberately created and spread by malicious actors. This paper provides a comprehensive survey of recent approaches to mitigating both types of misinformation in LLMs. It explores internal and external prevention methods, along with various techniques for misinformation tracing and detection. By evaluating the strengths and weaknesses of these approaches, this survey aims to illuminate the direction for future research in addressing misinformation in LLMs.

1 Introduction

In recent years, the rapid advancement of Large Language Models (LLMs) has led to the emergence of numerous outstanding works, including open-source projects such as LLaMA (Touvron et al., 2023) and Falcon (Almazrouei et al., 2023), as well as commercial products like GPT-4 (Achiam et al., 2023) and Gemini (Reid et al., 2024). These developments have facilitated a paradigm shift in the field of natural language processing, achieving significant progress not only in traditional NLP tasks such as question-answering systems (Tan et al., 2023), translation (Peng et al., 2023), and information extraction (Wei et al., 2023), but also demonstrating new capabilities in areas such as code generation (Poldrack et al., 2023).

Nevertheless, as LLMs are increasingly tested and applied across various domains, researchers have observed that their outputs are not always accurate. The erroneous outputs can be primarily categorized into two types: hallucinations and misinformation. Hallucinations refer to content generated by the model that includes entirely fictitious information (Zhang et al., 2023), which appears plausible but does not actually exist. Misinformation refers to content generated by the model that contains incorrect or inaccurate information, which may result from the model's misunderstanding of the problem or its reliance on incorrect training data (Chen and Shu, 2023). Therefore, misinformation is a broader concept compared to hallucinations. These erroneous outputs may undermine the utility of LLMs in critical applications and raise concerns about their responsible deployment in real-world scenarios.

Misinformation generated by LLMs can lead to serious consequences, such as manipulating public opinion, creating confusion, and spreading harmful ideologies. This issue is particularly concerning during sensitive periods, such as the COVID-19 pandemic, where misinformation exacerbated panic and misguidance (Zhou et al., 2023; Goldstein et al., 2023; Vykopal et al., 2023). The dangers posed by misinformation and hallucinations are diverse and have real-world impacts, underscoring the importance of addressing this challenge.

Misinformation generally falls into two categories: intentional misinformation and unintentional misinformation. Intentional misinformation often arises from malicious human manipulation of the model to generate specific erroneous information, while unintentional misinformation typically stems from data or model limitations, with hallucinations considered a subset of this category. Research indicates that hallucinations originate from various factors, including data quality issues, biases, outdated information, and limitations in model training and inference strategies (Rawte et al., 2023; Ji et al., 2023a). Intrinsic hallucinations arise from contradictions within the model's output, whereas extrinsic hallucinations manifest

as outputs that cannot be verified based on the input data.

Researchers have been diligently working to understand these causes and develop mitigation strategies, such as fact-oriented datasets (Thorne et al., 2018; Satapara et al., 2024), automatic data cleaning techniques (Li et al., 2024b), retrieval augmentation (Cai et al., 2022; Asai et al., 2023), and new model architectures and training objectives aimed at enhancing factual accuracy (He et al., 2023; Pan et al., 2023).

To counteract the spread of misinformation exacerbated by hallucinations, it is imperative to implement preventive measures throughout the entire lifecycle of LLMs and to develop methods for detecting whether certain information is misinformation generated by these models. Preventive measures include model internal and external methods, which affect the training inference stage and user input stage respectively. Detection approaches include watermark source tracking and factuality detecting, which first determine whether the information is generated by LLMs, and then determine if it is incorrect.

These comprehensive strategies aim to mitigate the adverse impact of misinformation and enhance the reliability and credibility of generated content. The primary contributions of this paper are as follows: 1). An exploration of the relationship and classification of hallucinations and misinformation; 2). A comprehensive survey of various methods for preventing and detecting misinformation; 3). A discussion of the limitations of current methods and suggestions for future research directions.

2 Misinformation

Misinformation is a significant challenge associated with LLMs. It can be broadly categorized into two types: unintentional misinformation, often resulting from hallucinations, and intentional misinformation, which is deliberately created and spread by malicious actors (Pan et al., 2023; Meyer and Choo, 2024; Hazzan, 2023).

2.1 Unintentional Misinformation

Hallucinations produced by LLMs constitute a significant source of misinformation (Galitsky, 2023; Quevedo et al., 2024; Nahar et al., 2024), which denote instances wherein LLMs produce information discordant with objective reality (Liu et al., 2024a; Andriopoulos and Pouwelse, 2023). Such occurrences typically arise when the model, without human intervention, endeavors to generate responses pertaining to topics it does not comprehensively grasp or possesses insufficient knowledge about (Ji et al., 2023b; McDonald et al., 2024; Tonmoy et al., 2024), resulting in Unintentional Misinformation. The etiology of hallucinations can be traced to several factors, including the incompleteness of training datasets (Yao et al., 2023), issues pertaining to data quality (Grover et al., 2024), and intrinsic limitations within the model's architecture (Xu et al., 2024). It is imperative to recognize that not all hallucinations are intentional; some may result from the model's misconstrual of information as predicated on its training (Grover et al., 2024; Lee et al., 2024).

To elaborate further, hallucinations can materialize in diverse forms, encompassing fabricated facts, fictitious entities, events, or statistics (Duan et al., 2024; Ji et al., 2023b; Yao et al., 2023). For instance, a LLM may erroneously assert the occurrence of a historical event in an incorrect year or attribute spurious quotations to eminent figures Certain hallucinations may be subtle, manifesting as minor inaccuracies, whereas others may constitute conspicuously glaring errors. Addressing the phenomenon of hallucinations is paramount for sustaining the credibility and utility of LLMs in pragmatic applications (Amatriain, 2024; Tonmoy et al., 2024; Ji et al., 2023b).

2.2 Intentional Misinformation

Intentional misinformation, also known as malicious attacks, is the deliberate creation and dissemination of false or misleading information by human actors (Matthews and Robertson, 2024). Unlike unintentional misinformation caused by hallucinations, intentional misinformation is not solely the result of technical or architectural limitations in the model (Ghai et al., 2024). It also involves the reliability and update frequency of data sources (Ruffo et al., 2023), as well as the context in which the model is deployed (Zhou et al., 2023).

Malicious actors may manipulate training data to mislead LLMs and perpetrate deceptive and misleading actions. The spread of malicious misinformation has more severe societal consequences, eroding trust in these models and their outputs among government entities and the general public (Murphy, 2022; Jamalzadeh, 2023; Williamson and Prybutok, 2024).

The ideal approach to dealing with misinformation is to prevent it from being generated by LLMs in the first place (Jimma, 2022; Huang et al., 2023a; Bodaghi et al., 2023). However, current technological methods cannot guarantee the complete prevention of hallucinations and misinformation (Li, 2023; Schlag et al., 2022). Therefore, prevention alone is not sufficient; it is also essential to consider detecting and verifying the output of large models to identify any potential misinformation (Di Sotto and Viviani, 2022; Shahid et al., 2022; Aïmeur et al., 2023).

Consequently, the subsequent chapters of this paper will focus primarily on methods for the prevention (§3) and detection (§4) of misinformation, the main process, classification logic, and method list of which can be shown in Figure 1. The correspondence between these methods and the model processes is discussed in Appendix A.

3 Prevention Strategies

In this section, we introduce how to prevent LLMs from generating misinformation. Based on whether modifying the internal state of the LLM is needed, we categorize the methods into two types: internal prevention (§3.1) and external prevention (§3.2).

3.1 Internal Prevention

Internal prevention methods prevent misinformation by adding information during training or using advanced decoding strategies during inference.

3.1.1 Adversarial Training

Adversarial training prevents LLMs from generating misinformation by making them more resilient to adversarial prompts—inputs specifically designed to trick the model into producing false information, which involves creating adversarial examples through small perturbations to the original data (Wang et al., 2019) and using these examples during training. By doing so, LLMs learn to better handle deceptive inputs and reduce misinformation output (Altinisik et al., 2022).

Research has demonstrated that adversarial training significantly enhances the ability of LLMs to withstand misinformation and evasion attacks (Chen and Shu, 2023; Al-Maliki et al., 2024). Additionally, it improves the model's security by preventing adversarial manipulations and unauthorized usage (Satapara et al., 2024).

3.1.2 Alignment Method

Alignment methods focus on aligning the LLM's outputs with human values and ethical guidelines. By fine-tuning models based on feedback from human reviewers, these methods aim to reduce harmful or misleading content, which can be categorized into the following classes:

Human-Preference Alignment. This method fine-tunes LLMs based on human feedback, ensuring that the generated content better reflects human values (Ji et al., 2024).

Self-Alignment. Although human-preference alignment improves value alignment, it relies heavily on external feedback. Self-alignment techniques enhance this by enabling LLMs to internally monitor and correct their outputs, thus improving resilience against adversarial manipulation without constant human oversight (Pang et al., 2024; Helbling et al., 2023; Xie et al., 2023).

Security and Privacy-Based Alignment. While self-alignment strengthens internal monitoring, it may not fully address security and privacy concerns. Security and privacy-based alignment methods ensure that models adhere to stringent security and privacy guidelines, thereby mitigating risks associated with data breaches and misuse (Yao et al., 2024; Liu et al., 2023d; Wang et al., 2024b).

Adversarial Attack Countermeasures. Even with security and privacy alignment, LLMs can still be vulnerable to sophisticated adversarial attacks. Adversarial attack countermeasures develop robust defenses against these attacks, ensuring that even well-aligned models maintain their integrity under targeted manipulation (Shah et al., 2023; Zou et al., 2023; Wang and Shu, 2023).

3.1.3 Decoding Method

Decoding methods involve techniques such as constrained decoding or selective sampling to ensure the generated text adheres to factual accuracy. These methods help prevent the generation of misleading or incorrect information.

Bi et al. (2024b) revisits factuality decoding methods, confirming their effectiveness in enhancing factual accuracy but warning they might hinder knowledge updates by making models overly confident in known facts, while Malon and Zhu (2024) discusses self-consistent decoding through learning from factuality preference rankings.

Jin et al. (2024a) proposes collaborative decoding, emphasizing critical tokens during the process to enhance the factuality of language models.

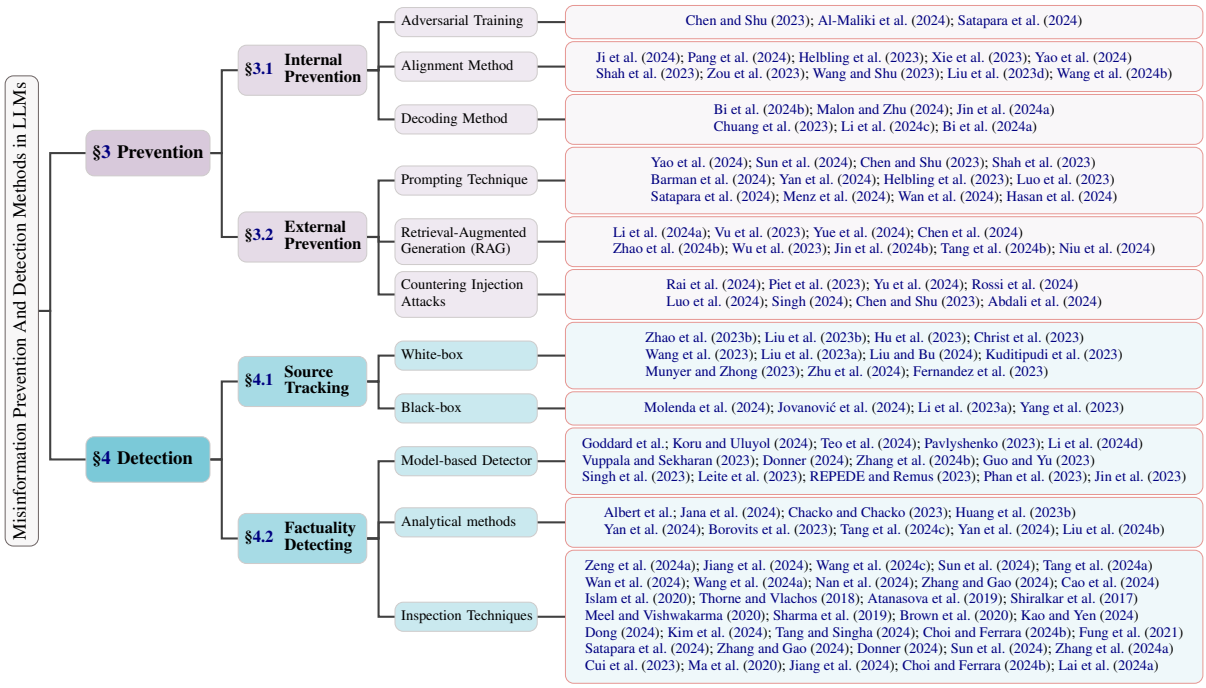

Figure 1: The main content flow and categorization of this survey.

Chuang et al. (2023) introduces Dola, a method that contrasts logits from different layers to emphasize high-level semantic information and weakon low-level grammatical information without extra fine-tuning.

Li et al. (2024c) presents Nearest Neighbor Speculative Decoding (NEST), which uses nearest neighbor matches to refine outputs, balancing speed and fluency while addressing hallucinations and providing attribution. Bi et al. (2024a) enhances model confidence in edited facts through knowledge contrasting decoding, showing significant improvements in factual accuracy.

3.2 External Prevention

This approach primarily targets the user's input, focusing on optimizing it to facilitate the model's reasoning and generation of factually accurate outputs, rather than modifying the model itself.

3.2.1 Prompting Technique

Prompting techniques involve carefully designed input prompts that guide the LLM to generate accurate and relevant information. Effective prompting can help steer the model away from producing misinformation.

Predefined Prompts focuses on using predefined prompts to guide LLMs in generating accurate and non-misleading content. Yao et al. (2024) discusses continuous prompts for secure code generation, Sun et al. (2024) explores predefined strategies for fake news, and Chen and Shu (2023) investigates using predefined prompts to reduce misinformation generation.

Persona-based Prompts involve having LLMs take on specific roles to generate content aligned with the persona's knowledge and characteristics. Shah et al. (2023) studies persona modulation for preventing jailbreaks, Barman et al. (2024) examines LLMs' roles in multimedia disinformation with persona prompts, and Yan et al. (2024) enhances rumor detection using persona and knowledge-powered prompts.

Adversarial Prompts are used to test and strengthen LLM robustness, ensuring accuracy even when faced with malicious prompts. Satapara et al. (2024) proposes adversarial prompting for generating a misinformation detection dataset, and Menz et al. (2024) investigates health misinformation prevention with adversarial techniques.

Knowledge-Powered Prompts incorporate domain-specific knowledge to enhance the accuracy and reliability of LLM outputs. Yan et al. (2024) and Luo et al. (2023) highlight improvements in rumor and hallucination detection, while Wan et al. (2024) introduces knowledge-powered prompting for generating reactions and explanations.

Security and Ethical Prompts ensure that LLM-generated content adheres to ethical standards and

safety protocols. Menz et al. (2024) examines safeguards risk mitigation, and transparency measures for health misinformation, and Hasan et al. (2024) focuses on increasing jailbreak resistance through pruning methods.

3.2.2 Retrieval-Augmented Generation (RAG)

This method involves enhancing the LLM's outputs by integrating information retrieved from external databases or websites, aims to ground the generated content in factual data, thus reducing the risk of misinformation.

Factual Accuracy Enhancement via RAG is discussed through various methods such as domain-specific queries with private knowledge-base in Li et al. (2024a), search engine augmentation proposed by FreshLLMs in Vu et al. (2023), evidence-driven response generation to counter misinformation in Yue et al. (2024), and reinforcement retrieval leveraging fine-grained feedback for fact-checking in Zhang and Gao (2024).

Chen et al. (2024) and Zhao et al. (2024b) discuss benchmarking LLMs in RAG and factuality evaluation respectively, and Wu et al. (2023) introduces a hallucination corpus for developing trustworthy retrieval-augmented models. Jin et al. (2024b) explores the use of RAG combined with LLMs for personalized disease prediction and data preprocessing to prevent data leakage. Tang et al. (2024b) proposes Self-Retrieval, an end-to-end, LLM-driven architecture to achieve document generation and self-assessment. Niu et al. (2024) presents a framework called Self-Refinement-Enhanced Knowledge Graph Retrieval (Re-KGR) to augment the factuality of LLMs' responses with less retrieval efforts in the medical field.

3.2.3 Countering Injection Attacks

This approach develops unique protection mechanisms to prevent the model from being manipulated through prompt injection attacks, thereby ensures the security and reliability of LLM outputs and is particularly effective in defending against intentional misinformation.

(Li et al., 2023b) establish a benchmark to evaluate the robustness of LLMs against prompt injection attacks. Rai et al. (2024) introduces a multi-tiered defense architecture named Defensive Prompt Patch (DPP) to protect LLMs, and Piet et al. (2023) presents a task-specific finetuning approach under the facts that LLMs can only follow instructions once they have undergone instruction tuning. Yu et al. (2024) investigates jailbreak prompts and their exploitation of LLM vulnerabilities, while Rossi et al. (2024) categorizes different types of prompt injection attacks and proposes standard countermeasures.

More prevention methods are disscussed in Appendix B.

4 Detection Methods

Preventive methods for misinformation are not always successful, necessitating the need to detect the outputs of large language models. The detection process involves two stages: the first stage is tracing whether the information was generated by LLMs, denoted as Source Tracking (§4.1), and the second stage is detecting whether the model's output contains misinformation, designated as Factuality Detecting (§4.2).

4.1 Source Tracking

Source Tracking refers to detecting whether a given text is generated by an LLM. This is typically done using two methods: white-box detection and black-box detection.

4.1.1 White-box Detection

A typical example of white-box detection methods is LLM watermarking (Liu et al., 2023c). By introducing detectable features during text generation, LLM watermarking can reliably determine if a text was generated by an LLM.

Currently, mainstream LLM watermarking is introduced during the inference phase, either by modifying the distribution for generating the next tokens (Kirchenbauer et al., 2023) or altering the sampling process (Kuditipudi et al., 2023). The main optimization directions for LLM watermarking are as follows:

Lossless Watermarking aims to introduce watermarks without degrading the quality of the generated text. Techniques in this category include reweight-watermark method (Hu et al., 2023) and methods that fix the sampling seed in LLM sampling (Christ et al., 2023).

More Robust Watermarking. A robust LLM watermarking technique should remain detectable even after the text has been modified. Approaches in this direction include global unified watermark (Zhao et al., 2023a), semantically invariant watermark (Liu et al., 2023b), and the use of edit distance

error correction codes to enhance robustness during watermark detection (Kuditipudi et al., 2023).

Multi-bit Watermarking intends to detect multi-bit information from watermarked text. (Yoo et al., 2023) distribute different watermark information to different positions in the text, while (Qu et al., 2024) use error correction codes to achieve robust multi-bit watermarking.

However, watermark-based detection typically requires intervention during LLM text generation, which may not be feasible in some scenarios.

4.1.2 Black-Box Detection

Black-box detection refers to detecting any text without adding explicit features to the text generated by LLMs. There are typically two approaches: one involves training classifiers to distinguish between LLM-generated text and human text (Lai et al., 2024b; Abburi et al., 2023), but this method lacks interpretability; the other involves identifying characteristic features of LLM-generated text to distinguish it from human text in a more interpretable way (Mitchell et al., 2023; Bao et al., 2023). However, as LLMs improve, the effectiveness of this feature-based approach diminishes.

4.2 Factuality Detecting

Through introducing external models (§4.2.1), conducting systematic analysis of the information (§4.2.2), and employing various inspection strategies (§4.2.3), we can examine whether a given text output from LLMs constitutes misinformation.

4.2.1 Model-based Detector

Labeled Classifier leverages labeled datasets with or without misinformation to train a model, such as traditional machine learning methods and deep neural networks to identify misinformation.

Koru and Uluyol (2024) focuses on using BERT models to classify fake news in Turkish tweets. Teo et al. (2024) compares LLMs with traditional machine learning models for fake news detection. Pavlyshenko (2023) analyzes the use of fine-tuned LLMs for disinformation detection. Vuppala and Sekharan (2023) discusses a fine-tuned LLM specifically for detecting click-bait titles. Nguyen et al. (2023) investigates fine-tuning LLMs for detecting predatory content. Donner (2024) evaluates different API-based misinformation detection methods using LLMs. Zhang et al. (2024b) reviews the mitigation of misinformation and social media manipulation using a combination of supervised learning and LLMs. Guo and Yu (2023) introduces AuthentiGPT, which uses black-box LLMs for detecting machine-generated text. Stewart et al. (2023) evaluates the use of transfer learning in fake news detection.

Pattern Detector uses clustering algorithms to group texts and detect anomalous clusters that may contain misinformation, or employ other unsupervised learning algorithms to detect anomalous texts. Singh et al. (2023) introduces an unsupervised method for retrieving debunked narratives, Leite et al. (2023) uses credibility signals and weak supervision, and REPEDE and Remus (2023) compares various AI models, including clustering techniques, for fake news detection. Tang et al. (2024c) discusses detecting LLM-generated text without labeled datasets. Mitchell et al. (2023) presents a zero-shot detection method using probability curvature.

Graph Neural Network (GNN) aims to leverage the structural properties of graphs to enhance the accuracy and robustness of identifying false information. Li et al. (2024d) discusses the integration of LLMs with GNNs for evidence-aware fake news detection. Phan et al. (2023) surveys various GNN methods for fake news detection, highlighting their applications and effectiveness. Jin et al. (2023) provides a comprehensive survey on the interaction between LLMs and GNNs. Xu et al. (2022) compares LLM-based and non-LLM-based misinformation detectors, focusing on the benefits of using GNNs for evidence-aware detection.

4.2.2 Analytical methods

Statistical and Rule-based Filter compares outputs from multiple LLMs to detect inconsistencies, or establish specific rules, such as checking facts against known data, to identify potential misinformation. Albert et al.; Jana et al. (2024) discussed the evolution and comparison of LLMs versus rule-based systems, highlighting limitations and improvements. Chacko and Chacko (2023) explored the paradigm shift in deep learning with advanced statistical methods. Yan et al. (2024); Borovits et al. (2023), Tang et al. (2024c) introduced hybrid techniques combining rule-based and statistical methods for tasks such as rumor detection and anonymization. Liu et al. (2023c); Xiang et al. (2024) surveyed specific applications like text watermarking and misinformation benchmarks, providing comprehensive analyses. Prajapati et al. (2024) focused on detecting AI-generated

text using various methods. Sternfeld et al. (2024) discusses entity triplet extraction for factual consistency, Subramaniam et al. (2023) highlights rule-based heuristics in numeric data search. Xiang et al. (2024) evaluates both rule-based and LLM approaches in maternity care misinformation.

Heuristic Methods establish a set of rules or heuristics based on linguistic features and logical consistency to detect misinformation, or use natural language processing techniques to identify common misinformation patterns and language characteristics. Huang et al. (2023b) provides a comprehensive review of heuristic methods for LLM hallucination, Yan et al. (2024) enhances rumor detection capabilities, and Liu et al. (2024b) integrates rule-based aggregation in a trustworthy framework.

Data Augmentation and Feature Selection involves augmenting data and selecting hybrid features rather than rule-based or heuristic methods. Wan et al. (2024) introduces a hybrid model with LLMs for data augmentation and explanatory features, Lai et al. (2024a) utilizes LLMs for data augmentation in fake news detection, and Du et al. (2024) designs eight features for complex instructions and construct a comprehensive evaluation dataset from real-world scenarios.

4.2.3 Inspection Techniques

Crowdsourced Verifier combines human expert feedback with LLM outputs through interactive iterations to enhance detection capabilities, and utilize user reports to verify misinformation and train the model to improve its detection abilities.

Zeng et al. (2024a) discusses the combination of LLMs with crowdsourced feedback, Banerjee et al. explores self-generated feedback by LLMs, and Jiang et al. (2024) looks at the evolving challenge of disinformation detection by combining AI and human efforts. Wang et al. (2024c) examines effective verification of LLM labels through human collaboration, Uchendu et al. (2023) investigates if human collaboration enhances the accuracy of identifying deepfake texts, and Zeng et al. (2024b) proposes a human-in-the-loop strategy for identifying similar data points.

Sun et al. (2024) studies the deceptive power of LLM-generated fake news, Wan et al. (2024) introduces a framework for generating reactions and explanations, and Wang et al. (2024a) teaches LLMs to interpret multimodal misinformation through knowledge distillation. Nan et al. (2024) enhances fake news detection with generated comments, Zhang and Gao (2024) uses fine-grained feedback for fact-checking, and Cao et al. (2024) evaluates the capability of LLMs in detecting misinformation in scientific news reporting.

Fact-checking uses external knowledge bases to automatically verify the authenticity of generated content, or constructs and uses knowledge graphs to fact-check and ensures generated statements align with known facts.

Automated Fact-Checking Methods leverage machine learning and NLP methods to automatically identify and verify the truthfulness of information. Islam et al. (2020) discusses deep learning approaches, while Thorne and Vlachos (2018) provides a comprehensive overview of task formulations and future directions. Atanasova et al. (2019) focuses on integrating factual and contextual relevance in automated fact-checking. Choi and Ferrara (2024a) describes automated claim matching for fact-checking. Brown et al. (2020) describes the capabilities of GPT-3 in fact-checking. Dong (2024); Kim et al. (2024) utilize multiple LLM-driven agents for fact-checking online discussions, generating reactions and explanations, and investigating the ability of LLMs to produce faithful explanations.

Compared to automated fact-checking methods with no external knowledge required, Knowledge-based Fact-Checking Methods rely on pre-constructed databases, specific domains or knowledge graphs to verify facts. Shiralkar et al. (2017) leverages knowledge graphs for real-time fact-checking. (Donner, 2024; Sun et al., 2024) evaluate various LLM-based methods for automated fact-checking within specific domains and using specialized datasets, such as the LIAR dataset. They focus on optimizing LLMs for accuracy and efficiency in domain-specific contexts. Meel and Vishwakarma (2020) surveys the opportunities and challenges in social media fact-checking, Ma et al. (2020) presents a data mining perspective on social media fake news detecting.

However, neither of the aforementioned methods utilize user feedback. Game-Based and Interactive Fact-Checking Tang and Singha (2024); Choi and Ferrara (2024b) leverage LLM-enhanced games and claim matching processes to study the effectiveness of interactive and engaging methods for fact-checking. These methods aim to increase user participation and accuracy in identifying misinfor-

mation. Additionally, there are other methods that place a greater emphasis on efficiency and optimization. (Tang et al., 2024a; Kao and Yen, 2024) propose efficient methods for fact-checking LLM-generated content and uses small models and self-refinement respectively to enhance domain generalization in automatic fact-checking. The focus is on improving processing efficiency and accuracy.

Consistency Checking generates or verifies the same information using multiple different models to check for consistency, or compare multiple outputs from the same model to ensure internal consistency of the information. Cui et al. (2023) introduces methods for enhancing the consistency of large language model outputs using a divide-and-conquer reasoning approach. Fung et al. (2021) proposes a cross-media approach for detecting fake news through fine-grained information consistency checking. Zhang et al. (2024a) explores efficient methods for identifying non-factual content using probe training combined with offline consistency checking. Jiang et al. (2024) examines Chain-of-Thought and Self-Consistency methods for detecting misinformation generated by large language models, highlighting the evolving challenges in this field. Choi and Ferrara (2024b) investigates the enhancement of fact-checking processes through claim matching with LLMs, emphasizing global consistency checks. Lai et al. (2024a) discusses the application of text consistency methods to improve fake news detection, using a large language model-based approach for data augmentation in rumor detection.

5 Open Problems and Future Directions

For prevention methods. Adversarial training and alignment methods in internal approaches (§3.1) may lead to overly conservative models, high bias risk, and vulnerability to attacks, while decoding methods might affect output diversity. In external methods (§3.2), prompting techniques and RAG rely on the quality of knowledge bases and retrieval system accuracy.

Future research directions include developing more efficient adversarial training techniques to reduce computational costs and enhance model flexibility, creating dynamic alignment methods to adjust to different users and contexts, and exploring smarter decoding strategies to balance consistency and diversity in responses. Additionally, adaptive prompting techniques that can adjust in real-time, enhancing the accuracy and comprehensiveness of retrieval-augmented generation (RAG), and building systematic defense frameworks against injection attacks are crucial.

For detection methods. White-box methods in Source Tracing (§4.1) face privacy and security concerns, complexity problems, and high computational costs. Future solutions include privacy-preserving techniques, simplifying detection processes, and automated tools. While black-box methods suffer from insufficient robustness, detection errors, and susceptibility to attacks. Potential solutions involve more robust watermark embedding, improved detection algorithms, and enhanced resistance to adversarial attacks.

In Factuality Detecting methods (§4.2), model-based detectors face challenges such as high costs of labeled data acquisition, limited generalization capabilities, and computational complexity. Future research directions include developing efficient data labeling methods, improving the accuracy of unsupervised and zero-shot methods, and optimizing GNN computational efficiency. Analytical methods primarily encounter issues with the adaptability of rule-based systems and the high data demands of statistical methods. Future studies should focus on creating adaptive rule systems and enhancing the performance of statistical methods on small datasets. Inspection techniques face challenges related to the quality of participant contributions, knowledge base updates, and high computational costs. Future directions involve improving participant screening and training and developing dynamically updated knowledge bases.

6 Conclusion

Misinformation generated by LLMs poses significant challenges to society, requiring concerted efforts from researchers, industry stakeholders, and policymakers. In this paper, we reviewed the categorization and sources of misinformation, highlighting the complexity and diversity of its manifestations. We examined various prevention methods, including internal and external approaches, and demonstrated a comprehensive detection methodology that includes two critical steps, source tracing and factuality detecting, which contribute to curbing the spread of misinformation. We then discussed potential issues present in current methods and proposed solutions to address these challenges, thereby providing direction for future research.

Limitations

This paper is subject to several limitations. The investigation into misinformation within LLMs is principally focused on the period spanning 2023-2024, a timeframe selected due to the explosive growth in LLM development. Furthermore, the scope of the survey is predominantly concentrated on methodologies associated with LLMs. However, the prevention and detection of misinformation also encompass various traditional approaches that, owing to their limited applicability to LLMs, were not expounded upon in this survey.

While it was our intention to comprehensively cover every aspect of misinformation prevention and detection in LLMs, constraints related to paper length necessitated a more succinct discussion of certain methodologies. In addition, the analysis of our survey results is qualitative rather than quantitative, lacking a detailed examination of the specific efficacy of each method. Hopefully, future researchers will further this line of inquiry.

References

Harika Abburi, Michael Suesserman, Nirmala Pudota, Balaji Veeramani, Edward Bowen, and Sanmitra Bhattacharya. 2023. Generative ai text classification using ensemble llm approaches. *arXiv preprint arXiv:2309.07755*.

Sara Abdali, Richard Anarfi, CJ Barberan, and Jia He. 2024. Securing large language models: Threats, vulnerabilities and responsible practices. *arXiv preprint arXiv:2403.12503*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf, and Muhammad Ovais Ahmad. 2020. Fake news detection using machine learning ensemble methods. *Complexity*, 2020:1–11.

Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30.

Shawqi Al-Maliki, Adnan Qayyum, Hassan Ali, Mohamed Abdallah, Junaid Qadir, Dinh Thai Hoang, Dusit Niyato, and Ala Al-Fuqaha. 2024. Adversarial machine learning for social good: Reframing the adversary as an ally. *IEEE Transactions on Artificial Intelligence*.

Julien Albert, Martin Balfroid, Miriam Doh, Jeremie Bogaert, Luca La Fisca, Liesbet De Vos, Bryan Renard, Vincent Stragier, and Emmanuel Jean. User preferences for large language model versus template-based explanations of movie recommendations: A pilot study.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Enes Altinisik, Hassan Sajjad, Husrev Taha Sencar, Safa Messaoud, and Sanjay Chawla. 2022. Impact of adversarial training on robustness and generalizability of language models. *arXiv preprint arXiv:2211.05523*.

Xavier Amatriain. 2024. Measuring and mitigating hallucinations in large language models: Amultifaceted approach.

Konstantinos Andriopoulos and Johan Pouwelse. 2023. Augmenting llms with knowledge: A survey on hallucination prevention. *arXiv preprint arXiv:2309.16459*.

Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. 2023. Tutorial proposal: Retrieval-based language models and applications. In *The 61st Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, page 41.

Pepa Atanasova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Georgi Karadzhov, Tsvetomila Mihaylova, Mitra Mohtarami, and James Glass. 2019. Automatic fact-checking using context and discourse information. *Journal of Data and Information Quality (JDIQ)*, 11(3):1–27.

Tanushree Banerjee, Richard Zhu, Runzhe Yang, Denis Peskov, Brandon Stewart, and Karthik Narasimhan. Llms are superior feedback providers: Bootstrapping reasoning for lie detection with self-generated feedback.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2023. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *arXiv preprint arXiv:2310.05130*.

Dipto Barman, Ziyi Guo, and Owen Conlan. 2024. The dark side of language models: Exploring the potential of llms in multimedia disinformation generation and dissemination. *Machine Learning with Applications*, page 100545.

Baolong Bi, Shenghua Liu, Lingrui Mei, Yiwei Wang, Pengliang Ji, and Xueqi Cheng. 2024a. Decoding by contrasting knowledge: Enhancing llms' confidence on edited facts. *arXiv preprint arXiv:2405.11613*.

Baolong Bi, Shenghua Liu, Yiwei Wang, Lingrui Mei, and Xueqi Cheng. 2024b. Is factuality decoding a free lunch for llms? evaluation on knowledge editing benchmark. *arXiv preprint arXiv:2404.00216*.

Arezo Bodaghi, Ketra A Schmitt, Pierre Watine, and Benjamin CM Fung. 2023. A literature review on detecting, verifying, and mitigating online misinformation. *IEEE Transactions on Computational Social Systems*.

Nemania Borovits, Gianluigi Bardelloni, Damian Andrew Tamburri, and Willem-Jan Van Den Heuvel. 2023. Anonymization-as-a-service: The service center transcripts industrial case. In *International Conference on Service-Oriented Computing*, pages 261–275. Springer.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Deng Cai, Yan Wang, Lemao Liu, and Shuming Shi. 2022. Recent advances in retrieval-augmented text generation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 3417–3419.

Yupeng Cao, Aishwarya Muralidharan Nair, Elyon Eyimife, Nastaran Jamalipour Soofi, KP Subbalakshmi, John R Wullert II, Chumki Basu, and David Shallcross. 2024. Can large language models detect misinformation in scientific news reporting? *arXiv preprint arXiv:2402.14268*.

Neha Chacko and Viju Chacko. 2023. Paradigm shift presented by large language models (llm) in deep learning. *ADVANCES IN EMERGING COMPUTING TECHNOLOGIES*, 40.

Canyu Chen and Kai Shu. 2023. Combating misinformation in the age of llms: Opportunities and challenges. *arXiv preprint arXiv:2311.05656*.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

Eun Cheol Choi and Emilio Ferrara. 2024a. Automated claim matching with large language models: empowering fact-checkers in the fight against misinformation. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1441–1449.

Eun Cheol Choi and Emilio Ferrara. 2024b. Fact-gpt: Fact-checking augmentation via claim matching with llms. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 883–886.

Miranda Christ, Sam Gunn, and Or Zamir. 2023. Undetectable watermarks for language models. *arXiv preprint arXiv:2306.09194*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Wendi Cui, Jiaxin Zhang, Zhuohang Li, Damien Lopez, Kamalika Das, Bradley Malin, and Sricharan Kumar. 2023. Evaluating and improving generation consistency of large language models via a divide-conquer-reasoning approach.

Stefano Di Sotto and Marco Viviani. 2022. Health misinformation detection in the social web: an overview and a data science approach. *International Journal of Environmental Research and Public Health*, 19(4):2173.

Yihan Dong. 2024. The multi-agent system based on llm for online discussions. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pages 2731–2733.

Catherine Donner. 2024. Misinformation detection methods using large language models and evaluation of application programming interfaces.

Yanrui Du, Sendong Zhao, Danyang Zhao, Ming Ma, Yuhan Chen, Liangyu Huo, Qing Yang, Dongliang Xu, and Bing Qin. 2024. Mogu: A framework for enhancing safety of open-sourced llms while preserving their usability. *arXiv preprint arXiv:2405.14488*.

Hanyu Duan, Yi Yang, and Kar Yan Tam. 2024. Do llms know about hallucination? an empirical investigation of llm's hidden states. *arXiv preprint arXiv:2402.09733*.

Yongkai Fan, Binyuan Xu, Linlin Zhang, Jinbao Song, Albert Zomaya, and Kuan-Ching Li. 2023. Validating the integrity of convolutional neural network predictions based on zero-knowledge proof. *Information Sciences*, 625:125–140.

Pierre Fernandez, Antoine Chaffin, Karim Tit, Vivien Chappelier, and Teddy Furon. 2023. Three bricks to consolidate watermarks for large language models. In *2023 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE.

Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avirup Sil. 2021. Infosurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1683–1698.

Boris A Galitsky. 2023. Truth-o-meter: Collaborating with llm in fighting its hallucinations.

Ambica Ghai, Pradeep Kumar, and Samrat Gupta. 2024. A deep-learning-based image forgery detection framework for controlling the spread of misinformation. *Information Technology & People*, 37(2):966–997.

Shlok Gilda. 2017. Notice of violation of ieee publication principles: Evaluating machine learning algorithms for fake news detection. In *2017 IEEE 15th student conference on research and development (SCOReD)*, pages 110–115. IEEE.

Jamal Goddard, Yuksel Celik, and Sanjay Goel. Beyond the human eye: Comprehensive approaches to ai text detection.

Josh A Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246*.

Shresth Grover, Vibhav Vineet, and Yogesh S Rawat. 2024. Navigating hallucinations for reasoning of unintentional activities. *arXiv preprint arXiv:2402.19405*.

Zhen Guo and Shangdi Yu. 2023. Authentigpt: Detecting machine-generated text via black-box language models denoising. *arXiv preprint arXiv:2311.07700*.

Joschka Haltaufderheide and Robert Ranisch. 2024. The ethics of chatgpt in medicine and healthcare: A systematic review on large language models (llms). *arXiv preprint arXiv:2403.14473*.

Asimul Haque and Muhammad Abulaish. 2022. A graph-based approach leveraging posts and reactions for detecting rumors on online social media. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 533–544.

Adib Hasan, Ileana Rugina, and Alex Wang. 2024. Pruning for protection: Increasing jailbreak resistance in aligned llms without fine-tuning. *arXiv preprint arXiv:2401.10862*.

Moses K Hazzan. 2023. Deception in the era of digital technologies and the distortion of reality and facts: An x-ray of nigerian peculiarities. *E-Learning and Digital Media*, 20(6):563–578.

Qianyu He, Jie Zeng, Wenhao Huang, Lina Chen, Jin Xiao, Qianxi He, Xunzhe Zhou, Lida Chen, Xintao Wang, Yuncheng Huang, et al. 2023. Can large language models understand real-world complex instructions? *arXiv preprint arXiv:2309.09150*.

Alec Helbling, Mansi Phute, Matthew Hull, and Duen Horng Chau. 2023. Llm self defense: By self examination, llms know they are being tricked. *arXiv preprint arXiv:2308.07308*.

Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. 2023. Unbiased watermark for large language models. *arXiv preprint arXiv:2310.10669*.

Chin-Tser Huang, Tieming Geng, and Jian Liu. 2023a. Capturing the characteristics of mis/disinformation propagation over the internet. In *Disruptive Technologies in Information Sciences VII*, volume 12542, pages 187–195. SPIE.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023b. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Yiting Huang, Yong Yu, Huilin Li, Yannan Li, and Aikui Tian. 2022. Blockchain-based continuous data integrity checking protocol with zero-knowledge privacy protection. *Digital Communications and Networks*, 8(5):604–613.

Md Rafiqul Islam, Shaowu Liu, Xianzhi Wang, and Guandong Xu. 2020. Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining*, 10(1):82.

Saeed Jamalzadeh. 2023. Protecting infrastructure networks from disinformation.

Sunit Jana, Rakhi Biswas, Koushik Pal, Suparna Biswas, and Kaushik Roy. 2024. The evolution and impact of large language model systems: A comprehensive analysis.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2024. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023a. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023b. Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843.

Bohan Jiang, Zhen Tan, Ayushi Nirmal, and Huan Liu. 2024. Disinformation detection: An evolving challenge in the age of llms. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*, pages 427–435. SIAM.

Ethiopia Jimma. 2022. *College of Law and Governance School of Law LLM in Human Rights and Criminal Law*. Ph.D. thesis, Jimma University.

Bowen Jin, Gang Liu, Chi Han, Meng Jiang, Heng Ji, and Jiawei Han. 2023. Large language models on graphs: A comprehensive survey. *arXiv preprint arXiv:2312.02783*.

Lifeng Jin, Baolin Peng, Linfeng Song, Haitao Mi, Ye Tian, and Dong Yu. 2024a. Collaborative decoding of critical tokens for boosting factuality of large language models. *arXiv preprint arXiv:2402.17982*.

Mingyu Jin, Qinkai Yu, Dong Shu, Chong Zhang, Lizhou Fan, Wenyue Hua, S Zhu, Y Meng, Z Wang, M Du, et al. 2024b. Health-llm: Personalized retrieval-augmented disease prediction system. *arXiv preprint arXiv:2402.00746*.

Nikola Jovanović, Robin Staab, and Martin Vechev. 2024. Watermark stealing in large language models. *arXiv preprint arXiv:2402.19361*.

Wei-Yu Kao and An-Zi Yen. 2024. Magic: Multi-argument generation with self-refinement for domain generalization in automatic fact-checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10891–10902.

Zeba Khanam, BN Alwasel, H Sirafi, and Mamoon Rashid. 2021. Fake news detection using machine learning approaches. In *IOP conference series: materials science and engineering*, volume 1099, page 012040. IOP Publishing.

Kyungha Kim, Sangyun Lee, Kung-Hsiang Huang, Hou Pong Chan, Manling Li, and Heng Ji. 2024. Can llms produce faithful explanations for fact-checking? towards faithful explainable fact-checking via multi-agent debate. *arXiv preprint arXiv:2402.07401*.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *International Conference on Machine Learning*, pages 17061–17084. PMLR.

Gülsüm Kayabaşi Koru and Çelebı Uluyol. 2024. Detection of turkish fake news from tweets with bert models. *IEEE Access*.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. 2023. Robust distortion-free watermarks for language models. *arXiv preprint arXiv:2307.15593*.

Oleksandr Kuznetsov, Alex Rusnak, Anton Yezhov, Dzianis Kanonik, Kateryna Kuznetsova, and Stanislav Karashchuk. 2024. Enhanced security and efficiency in blockchain with aggregated zero-knowledge proof mechanisms. *arXiv preprint arXiv:2402.03834*.

Jianqiao Lai, Xinran Yang, Wenyue Luo, Linjiang Zhou, Langchen Li, Yongqi Wang, and Xiaochuan Shi. 2024a. Rumorllm: A rumor large language model-based fake-news-detection data-augmentation approach. *Applied Sciences*, 14(8):3532.

Zhixin Lai, Xuesheng Zhang, and Suiyao Chen. 2024b. Adaptive ensembles of fine-tuned transformers for llm-generated text detection. *arXiv preprint arXiv:2403.13335*.

Dongyub Lee, Eunhwan Park, Hodong Lee, and Heui-Seok Lim. 2024. Ask, assess, and refine: Rectifying factual consistency and hallucination in llms with metric-guided feedback learning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2422–2433.

João A Leite, Olesya Razuvayevskaya, Kalina Bontcheva, and Carolina Scarton. 2023. Detecting misinformation with llm-predicted credibility signals and weak supervision. *arXiv preprint arXiv:2309.07601*.

Jiarui Li, Ye Yuan, and Zehua Zhang. 2024a. Enhancing llm factual accuracy with rag to counter hallucinations: A case study on domain-specific queries in private knowledge-bases. *arXiv preprint arXiv:2403.10446*.

Ming Li, Lichang Chen, Jiuhai Chen, Shwai He, Jiuxiang Gu, and Tianyi Zhou. 2024b. Selective reflection-tuning: Student-selected data recycling for llm instruction-tuning. *arXiv preprint arXiv:2402.10110*.

Minghan Li, Xilun Chen, Ari Holtzman, Beidi Chen, Jimmy Lin, Wen-tau Yih, and Xi Victoria Lin. 2024c. Nearest neighbor speculative decoding for llm generation and attribution. *arXiv preprint arXiv:2405.19325*.

Peixuan Li, Pengzhou Cheng, Fangqi Li, Wei Du, Haodong Zhao, and Gongshen Liu. 2023a. Plmmark: a secure and robust black-box watermarking framework for pre-trained language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14991–14999.

Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. 2024d. Large language model agent for fake news detection. *arXiv preprint arXiv:2405.01593*.

Zekun Li, Baolin Peng, Pengcheng He, and Xifeng Yan. 2023b. Do you really follow me? adversarial instructions for evaluating the robustness of large language models. *arXiv preprint arXiv:2308.10819*.

Zihao Li. 2023. The dark side of chatgpt: legal and ethical challenges from stochastic parrots and hallucination. *arXiv preprint arXiv:2304.14347*.

Aiwei Liu, Leyi Pan, Xuming Hu, Shu'ang Li, Lijie Wen, Irwin King, and Philip S Yu. 2023a. A private watermark for large language models. *arXiv preprint arXiv:2307.16230*.

Aiwei Liu, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen. 2023b. A semantic invariant robust watermark for large language models. *arXiv preprint arXiv:2310.06356*.

Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Lijie Wen, Irwin King, and Philip S Yu. 2023c. A survey of text watermarking in the era of large language models. *arXiv preprint arXiv:2312.07913*.

Fang Liu, Yang Liu, Lin Shi, Houkun Huang, Ruifeng Wang, Zhen Yang, and Li Zhang. 2024a. Exploring and evaluating hallucinations in llm-powered code generation. *arXiv preprint arXiv:2404.00971*.

Hui Liu, Wenya Wang, Haoru Li, and Haoliang Li. 2024b. Teller: A trustworthy framework for explainable, generalizable and controllable fake news detection. *arXiv preprint arXiv:2402.07776*.

Tianrui Liu, Qi Cai, Changxin Xu, Zhanxin Zhou, Fanghao Ni, Yuxin Qiao, and Tsungwei Yang. 2024c. Rumor detection with a novel graph neural network approach. *arXiv preprint arXiv:2403.16206*.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023d. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

Yepeng Liu and Yuheng Bu. 2024. Adaptive text watermark for large language models.

Junyu Luo, Cao Xiao, and Fenglong Ma. 2023. Zero-resource hallucination prevention for large language models. *arXiv preprint arXiv:2309.02654*.

Yifeng Luo, Yupeng Li, Dacheng Wen, and Liang Lan. 2024. Message injection attack on rumor detection under the black-box evasion setting using large language model. In *Proceedings of the ACM on Web Conference 2024*, pages 4512–4522.

L Ma, W Gao, Z Wei, and Q Lu. 2020. Detecting fake news on social media: A data mining perspective. *IEEE Access*, 8:118121–118140.

Christopher Malon and Xiaodan Zhu. 2024. Self-consistent decoding for more factual open responses. *arXiv preprint arXiv:2403.00696*.

Luke J Matthews and Paul Robertson. 2024. *Theorizing the anthropology of belief: Magic, conspiracies, and misinformation*. Taylor & Francis.

Daniel McDonald, Rachael Papadopoulos, and Leslie Benningfield. 2024. Reducing llm hallucination using knowledge distillation: A case study with mistral large and mmlu benchmark. *Authorea Preprints*.

Priyanka Meel and Dinesh Kumar Vishwakarma. 2020. Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Systems with Applications*, 153:112986.

Bradley D Menz, Nicole M Kuderer, Stephen Bacchi, Natansh D Modi, Benjamin Chin-Yee, Tiancheng Hu, Ceara Rickard, Mark Haseloff, Agnes Vitry, Ross A McKinnon, et al. 2024. Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation: repeated cross sectional analysis. *bmj*, 384.

Marco Meyer and Chun Wei Choo. 2024. Harming by deceit: Epistemic malevolence and organizational wrongdoing. *Journal of Business Ethics*, 189(3):439–452.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, pages 24950–24962. PMLR.

Piotr Molenda, Adian Liusie, and Mark JF Gales. 2024. Waterjudge: Quality-detection trade-off when watermarking large language models. *arXiv preprint arXiv:2403.19548*.

Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M Bronstein. 2019. Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*.

Travis Munyer and Xin Zhong. 2023. Deeptextmark: Deep learning based text watermarking for detection of large language model generated text. *arXiv preprint arXiv:2305.05773*.

Brian J Murphy. 2022. *The Impact of Social Media Conveyed Russian-Backed Disinformation in a Polarized America: An Examination of the Executive Branch's Ethical Responsibility to Respond*. Ph.D. thesis, Georgetown University.

Mahjabin Nahar, Haeseung Seo, Eun-Ju Lee, Aiping Xiong, and Dongwon Lee. 2024. Fakes of varying shades: How warning affects human perception and engagement regarding llm hallucinations. *arXiv preprint arXiv:2404.03745*.

Qiong Nan, Qiang Sheng, Juan Cao, Beizhe Hu, Danding Wang, and Jintao Li. 2024. Let silence speak: Enhancing fake news detection with generated comments from large language models. *arXiv preprint arXiv:2405.16631*.

Maziar Nekovee, Yamir Moreno, Ginestra Bianconi, and Matteo Marsili. 2007. Theory of rumour spreading in complex social networks. *Physica A: Statistical Mechanics and its Applications*, 374(1):457–470.

Thanh Thi Nguyen, Campbell Wilson, and Janis Dalins. 2023. Fine-tuning llama 2 large language models for detecting online sexual predatory chats and abusive texts. *arXiv preprint arXiv:2308.14683*.

Mengjia Niu, Hao Li, Jie Shi, Hamed Haddadi, and Fan Mo. 2024. Mitigating hallucinations in large language models via self-refinement-enhanced knowledge retrieval. *arXiv preprint arXiv:2405.06545*.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2018. A survey on natural language processing for fake news detection. *arXiv preprint arXiv:1811.00770*.

Aishika Pal, Moumita Pradhan, et al. 2023. Survey of fake news detection using machine intelligence approach. *Data & Knowledge Engineering*, 144:102118.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. 2023. On the risk of misinformation pollution with large language models. *arXiv preprint arXiv:2305.13661*.

Xianghe Pang, Shuo Tang, Rui Ye, Yuxin Xiong, Bolun Zhang, Yanfeng Wang, and Siheng Chen. 2024. Self-alignment of large language models via multi-agent social simulation. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Bohdan M Pavlyshenko. 2023. Analysis of disinformation and fake news detection using fine-tuned large language model. *arXiv preprint arXiv:2309.04704*.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of chatgpt for machine translation. *arXiv preprint arXiv:2303.13780*.

Huyen Trang Phan, Ngoc Thanh Nguyen, and Dosam Hwang. 2023. Fake news detection: A survey of graph neural network methods. *Applied Soft Computing*, page 110235.

Julien Piet, Maha Alrashed, Chawin Sitawarin, Sizhe Chen, Zeming Wei, Elizabeth Sun, Basel Alomair, and David Wagner. 2023. Jatmo: Prompt injection defense by task-specific finetuning. *arXiv preprint arXiv:2312.17673*.

Andrés Piñeiro-Martín, Carmen García-Mateo, Laura Docío-Fernández, and María del Carmen López-Pérez. 2023. Ethical challenges in the development of virtual assistants powered by large language models. *Electronics*, 12(14):3170.

Russell A Poldrack, Thomas Lu, and Gašper Beguš. 2023. Ai-assisted coding: Experiments with gpt-4. *arXiv preprint arXiv:2304.13187*.

Manish Prajapati, Santos Kumar Baliarsingh, Chinmayee Dora, Ashutosh Bhoi, Jhalak Hota, and Jasaswi Prasad Mohanty. 2024. Detection of ai-generated text using large language model. In *2024 International Conference on Emerging Systems and Intelligent Computing (ESIC)*, pages 735–740. IEEE.

Wenjie Qu, Dong Yin, Zixin He, Wei Zou, Tianyang Tao, Jinyuan Jia, and Jiaheng Zhang. 2024. Provably robust multi-bit watermarking for ai-generated text via error correction code. *arXiv preprint arXiv:2401.16820*.

Ernesto Quevedo, Jorge Yero, Rachel Koerner, Pablo Rivas, and Tomas Cerny. 2024. Detecting hallucinations in large language model generation: A token probability approach. *arXiv preprint arXiv:2405.19648*.

Parijat Rai, Saumil Sood, Vijay K Madisetti, and Arshdeep Bahga. 2024. Guardian: A multi-tiered defense architecture for thwarting prompt injection attacks on llms. *Journal of Software Engineering and Applications*, 17(1):43–68.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Ștefan Emil REPEDE and BRAD Remus. 2023. A comparison of artificial intelligence models used for fake news detection. *BULLETIN OF" CAROL I" NATIONAL DEFENCE UNIVERSITY*, 12(1):114–131.

Sippo Rossi, Alisia Marianne Michel, Raghava Rao Mukkamala, and Jason Bennett Thatcher. 2024. An early categorization of prompt injection attacks on large language models. *arXiv preprint arXiv:2402.00898*.

Giancarlo Ruffo, Alfonso Semeraro, Anastasia Giachanou, and Paolo Rosso. 2023. Studying fake news spreading, polarisation dynamics, and manipulation by bots: A tale of networks and language. *Computer science review*, 47:100531.

Somya Ranjan Sahoo and Brij B Gupta. 2021. Multiple features based approach for automatic fake news detection on social networks using deep learning. *Applied Soft Computing*, 100:106983.

Shrey Satapara, Parth Mehta, Debasis Ganguly, and Sandip Modha. 2024. Fighting fire with fire: Adversarial prompting to generate a misinformation detection dataset. *arXiv preprint arXiv:2401.04481*.

Anne K Schlag, Jacob Aday, Iram Salam, Jo C Neill, and David J Nutt. 2022. Adverse effects of psychedelics: From anecdotes and misinformation to systematic science. *Journal of Psychopharmacology*, 36(3):258–272.

Rusheb Shah, Soroush Pour, Arush Tagade, Stephen Casper, Javier Rando, et al. 2023. Scalable and transferable black-box jailbreaks for language models via persona modulation. *arXiv preprint arXiv:2311.03348*.

Wajiha Shahid, Bahman Jamshidi, Saqib Hakak, Haruna Isah, Wazir Zada Khan, Muhammad Khurram Khan, and Kim-Kwang Raymond Choo. 2022. Detecting

and mitigating the dissemination of fake news: Challenges and future research opportunities. *IEEE Transactions on Computational Social Systems*.

Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. 2019. Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(3):1–42.

Prashant Shiralkar, Alessandro Flammini, Filippo Menczer, and Giovanni Luca Ciampaglia. 2017. Finding streams in knowledge graphs to support fact checking. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 859–864. IEEE.

Iknoor Singh, Carolina Scarton, and Kalina Bontcheva. 2023. Utdrm: unsupervised method for training debunked-narrative retrieval models. *EPJ Data Science*, 12(1):59.

Shridhar Singh. 2024. Enhancing privacy and security in large-language models: A zero-knowledge proof approach. In *International Conference on Cyber Warfare and Security*, volume 19, pages 574–582.

Alexander Sternfeld, Andrei Kucharavy, Dimitri Percia David, Alain Mermoud, and Julian Jang-Jaccard. 2024. Llm-resilient bibliometrics: Factual consistency through entity triplet extraction.

Jake Stewart, Nikita Lyubashenko, and George Stefanek. 2023. The efficacy of detecting ai-generated fake news using transfer learning. *Issues in Information Systems*, 24(2).

Pranav Subramaniam, Udayan Khurana, Kavitha Srinivas, and Horst Samulowitz. 2023. Related table search for numeric data using large language models and enterprise knowledge graphs. In *ACM International Conference on Information and Knowledge Management*.

Xiaoqiang Sun, F Richard Yu, Peng Zhang, Zhiwei Sun, Weixin Xie, and Xiang Peng. 2021. A survey on zero-knowledge proof in blockchain. *IEEE network*, 35(4):198–205.

Yanshen Sun, Jianfeng He, Limeng Cui, Shuo Lei, and Chang-Tien Lu. 2024. Exploring the deceptive power of llm-generated fake news: A study of real-world detection challenges. *arXiv preprint arXiv:2403.18249*.

Mateusz Szczepański, Marek Pawlicki, Rafał Kozik, and Michał Choraś. 2021. New explainability method for bert-based model in fake news detection. *Scientific reports*, 11(1):23705.

Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023. Can chatgpt replace traditional kbqa models? an in-depth analysis of the question answering performance of the gpt llm family. In *International Semantic Web Conference*, pages 348–367. Springer.

Haoheng Tang and Mrinalini Singha. 2024. A mystery for you: A fact-checking game enhanced by large language models (llms) and a tangible interface. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–5.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024a. Minicheck: Efficient fact-checking of llms on grounding documents. *arXiv preprint arXiv:2404.10774*.

Qiaoyu Tang, Jiawei Chen, Bowen Yu, Yaojie Lu, Cheng Fu, Haiyang Yu, Hongyu Lin, Fei Huang, Ben He, Xianpei Han, et al. 2024b. Self-retrieval: Building an information retrieval system with one large language model. *arXiv preprint arXiv:2403.00801*.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2024c. The science of detecting llm-generated text. *Communications of the ACM*, 67(4):50–59.

Ting Wei Teo, Hui Na Chua, Muhammed Basheer Jasser, and Richard TK Wong. 2024. Integrating large language models and machine learning for fake news detection. In *2024 20th IEEE International Colloquium on Signal Processing & Its Applications (CSPA)*, pages 102–107. IEEE.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. *arXiv preprint arXiv:1806.07687*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Adaku Uchendu, Jooyoung Lee, Hua Shen, Thai Le, Dongwon Lee, et al. 2023. Does human collaboration enhance the accuracy of identifying llm-generated deepfake texts? In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 11, pages 163–174.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. 2023. Freshllms: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214*.

Pavan Sai Vuppala and Chandra N Sekharan. 2023. A fine-tuned large language model for improved clickbait title detection.

Ivan Vykopal, Matúš Pikuliak, Ivan Srba, Robert Moro, Dominik Macko, and Maria Bielikova. 2023. Disinformation capabilities of large language models. *arXiv preprint arXiv:2311.08838*.

Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. 2024. Dell: Generating reactions and explanations for llm-based misinformation detection. *arXiv preprint arXiv:2402.10426*.

Dilin Wang, Chengyue Gong, and Qiang Liu. 2019. Improving neural language modeling via adversarial training. In *International Conference on Machine Learning*, pages 6555–6565. PMLR.

Haizhou Wang, Sen Wang, and YuHu Han. 2022. Detecting fake news on chinese social media based on hybrid feature fusion method. *Expert Systems with Applications*, 208:118111.

Haoran Wang and Kai Shu. 2023. Backdoor activation attack: Attack large language models using activation steering for safety-alignment. *arXiv preprint arXiv:2311.09433*.

Lean Wang, Wenkai Yang, Deli Chen, Hao Zhou, Yankai Lin, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Towards codable text watermarking for large language models. *arXiv preprint arXiv:2307.15992*.

Longzheng Wang, Xiaohan Xu, Lei Zhang, Jiarui Lu, Yongxiu Xu, Hongbo Xu, and Chuang Zhang. 2024a. Mmidr: Teaching large language model to interpret multimodal misinformation via knowledge distillation. *arXiv preprint arXiv:2403.14171*.

Shenao Wang, Yanjie Zhao, Xinyi Hou, and Haoyu Wang. 2024b. Large language model supply chain: A research agenda. *arXiv preprint arXiv:2404.12736*.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024c. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2023. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*.

Steven M Williamson and Victor Prybutok. 2024. The era of artificial intelligence deception: Unraveling the complexities of false realities and emerging threats of misinformation. *Information*, 15(6):299.

Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Cheng Niu, Randy Zhong, Juntong Song, and Tong Zhang. 2023. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. *arXiv preprint arXiv:2401.00396*.

Tong Xiang, Liangzhi Li, Wangyue Li, Mingbai Bai, Lu Wei, Bowen Wang, and Noa Garcia. 2024. Caremi: chinese benchmark for misinformation evaluation in maternity and infant care. *Advances in Neural Information Processing Systems*, 36.

Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, Xing Xie, and Fangzhao Wu. 2023. Defending chatgpt against jailbreak attack via self-reminders. *Nature Machine Intelligence*, 5(12):1486–1496.

Weizhi Xu, Junfei Wu, Qiang Liu, Shu Wu, and Liang Wang. 2022. Evidence-aware fake news detection with graph neural networks. In *Proceedings of the ACM web conference 2022*, pages 2501–2510.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Yeqing Yan, Peng Zheng, and Yongjun Wang. 2024. Enhancing large language model capabilities for rumor detection with knowledge-powered prompting. *Engineering Applications of Artificial Intelligence*, 133:108259.

Xi Yang, Kejiang Chen, Weiming Zhang, Chang Liu, Yuang Qi, Jie Zhang, Han Fang, and Nenghai Yu. 2023. Watermarking text generated by black-box language models. *arXiv preprint arXiv:2305.08883*.

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, and Li Yuan. 2023. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211.

KiYoon Yoo, Wonhyuk Ahn, and Nojun Kwak. 2023. Advancing beyond identification: Multi-bit watermark for language models. *arXiv preprint arXiv:2308.00221*.

Zhiyuan Yu, Xiaogeng Liu, Shunning Liang, Zach Cameron, Chaowei Xiao, and Ning Zhang. 2024. Don't listen to me: Understanding and exploring jailbreak prompts of large language models. *arXiv preprint arXiv:2403.17336*.

Zhenrui Yue, Huimin Zeng, Yimeng Lu, Lanyu Shang, Yang Zhang, and Dong Wang. 2024. Evidence-driven retrieval augmented response generation for online misinformation. *arXiv preprint arXiv:2403.14952*.

Sergey Zapechnikov. 2020. Privacy-preserving machine learning as a tool for secure personalized information services. *Procedia Computer Science*, 169:393–399.

Xia Zeng, David La Barbera, Kevin Roitero, Arkaitz Zubiaga, Stefano Mizzaro, et al. 2024a. Combining large language models and crowdsourcing for hybrid human-ai misinformation detection.

Xianlong Zeng, Fanghao Song, and Ang Liu. 2024b. Similar data points identification with llm: A human-in-the-loop strategy using summarization and hidden state insights. *arXiv preprint arXiv:2404.04281*.

Xiaokang Zhang, Zijun Yao, Jing Zhang, Kaifeng Yun, Jifan Yu, Juanzi Li, and Jie Tang. 2024a. Transferable and efficient non-factual content detection via probe training with offline consistency checking. *arXiv preprint arXiv:2404.06742*.

Xuan Zhang and Wei Gao. 2024. Reinforcement retrieval leveraging fine-grained feedback for fact checking news claims with black-box llm. *arXiv preprint arXiv:2404.17283*.

Yizhou Zhang, Karishma Sharma, Lun Du, and Yan Liu. 2024b. Toward mitigating misinformation and social media manipulation in llm era. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1302–1305.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. 2023a. Provable robust watermarking for ai-generated text.

Xuandong Zhao, Yu-Xiang Wang, and Lei Li. 2023b. Protecting language generation models via invisible watermarking. In *International Conference on Machine Learning*, pages 42187–42199. PMLR.

Yanjie Zhao, Xinyi Hou, Shenao Wang, and Haoyu Wang. 2024a. Llm app store analysis: A vision and roadmap. *arXiv preprint arXiv:2404.12737*.

Yiran Zhao, Jinghan Zhang, I Chern, Siyang Gao, Pengfei Liu, Junxian He, et al. 2024b. Felm: Benchmarking factuality evaluation of large language models. *Advances in Neural Information Processing Systems*, 36.

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20.

Chaoyi Zhu, Jeroen Galjaard, Pin-Yu Chen, and Lydia Y Chen. 2024. Duwak: Dual watermarks in large language models. *arXiv preprint arXiv:2403.13000*.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

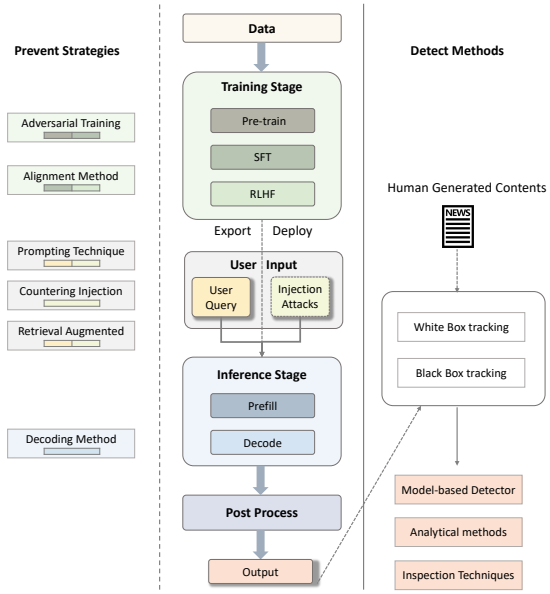

Figure 2: Correspondence diagram of misinformation prevention and detection methods throughout the entire lifecycle of LLMs

A Detailed Model Process Analysis

Figure 2 illustrates the correspondence between the prevention and detection methods within the entire LLM workflow. For the prevention method, the internal approach (§3.1) encompasses both the training and inference stages, while the external approach (§3.2) primarily pertains to the user input process. In the detection method, the Source tracking (§4.1) approach is focused on distinguishing between model outputs and human-generated content, whereas the Factuality detecting (§4.2) approach primarily targets misinformation detection in the outputs of LLMs.

B Additional Prevent Methods

B.1 Rumor Evaluation and Control

Implementing strategies to evaluate and control the output of LLMs to prevent the spread of rumors and misinformation. This includes techniques to assess the reliability of the generated content.

Machine Learning-Based Rumor Detection explores various machine learning models for detecting fake news. Pal et al. (2023) surveys different models, Monti et al. (2019) focuses on social media applications, and Gilda (2017) evaluates the performance of these models.

NLP-Based Rumor Detection covers the use of natural language processing techniques to identify fake news. Oshikawa et al. (2018) reviews NLP methods, Khanam et al. (2021) combines NLP with machine learning, and Szczepański et al. (2021) employs BERT-based models for detection.

Graph-Based Rumor Propagation Analysis investigates the use of graph analysis to understand and detect rumor propagation. Haque and Abulaish (2022) uses a graph-based approach, Liu et al. (2024c) employs graph neural networks, and Nekovee et al. (2007) applies graph theory for propagation analysis.

Hybrid and Multimodal Methods looks at combining multiple methods for detecting fake news. Wang et al. (2022) combines text and visual cues, and Ahmad et al. (2020) focuses on ensemble methods.

Social Media Data Analysis examines the analysis of social media data for fake news detection. Sahoo and Gupta (2021) studies user behavior on social platforms.

B.2 Zero-Knowledge Proof Approach

Using zero-knowledge proof methods to enhance privacy and security when handling sensitive information, thereby reducing the risk of information tampering and misrepresentation.

Data Privacy Protection explores the use of Zero-Knowledge Proofs (ZKP) in ensuring data privacy. Sun et al. (2021) discusses ZKP for secure data sharing, while Zapechnikov (2020) focuses on protecting user data in machine learning. addresses theoretical foundations and practical implementations of ZKP in data verification.

Authenticity Verification of Model Outputs covers the application of ZKP to verify the authenticity of AI-generated outputs. Fan et al. (2023) presents a novel approach to using ZKP for verifying AI content integrity. explores how ZKP can ensure trustworthy AI outputs, and Singh (2024) focuses on preventing misinformation in AI systems using ZKP.

Combining Blockchain Technology investigates the integration of ZKP with blockchain technology to enhance data security and prevent misinformation. Sun et al. (2021) discusses the combination of blockchain and ZKP for secure AI systems. Kuznetsov et al. (2024) examines the benefits of this combination for securing data, and Huang et al. (2022) explores blockchain-based ZKP for verifiable AI outputs.

B.3 Domain specific Misinformation

Health Information Misinformation Prevention focuses on measures to prevent health-related mis-

information using LLMs. Menz et al. (2024) evaluates risk mitigation and transparency measures against health disinformation. Piñeiro-Martín et al. (2023) discusses ethical challenges in developing LLM-powered virtual assistants, with an emphasis on transparency. Haltaufderheide and Ranisch (2024) systematically reviews ethical issues in medicine and healthcare related to LLMs, stressing patient privacy and data transparency.

Misinformation Prevention in Other Fields includes various methods to prevent misinformation in fields other than health. Zhao et al. (2024a) presents a vision for LLM app stores, addressing transparency and misinformation prevention. Zhou et al. (2023) examines AI-generated misinformation and both algorithmic and human solutions.