

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/375061964>

# Combating Misinformation in the Age of LLMs: Opportunities and Challenges

Preprint · October 2023

DOI: 10.13140/RG.2.2.28786.27844

CITATIONS

0

READS

681

2 authors:



Canyu Chen

Illinois Institute of Technology

14 PUBLICATIONS 10 CITATIONS

SEE PROFILE



Kai Shu

Illinois Institute of Technology

152 PUBLICATIONS 8,297 CITATIONS

SEE PROFILE

# Combating Misinformation in the Age of LLMs: Opportunities and Challenges

Canyu Chen  
Illinois Institute of Technology  
Chicago IL, USA  
cchen151@hawk.iit.edu

Kai Shu  
Illinois Institute of Technology  
Chicago IL, USA  
kshu@iit.edu

## Abstract

Misinformation such as fake news and rumors is a serious threat to information ecosystems and public trust. The emergence of Large Language Models (LLMs) has great potential to reshape the landscape of combating misinformation. Generally, LLMs can be a double-edged sword in the fight. On the one hand, LLMs bring promising opportunities for combating misinformation due to their profound world knowledge and strong reasoning abilities. Thus, one emergent question is: *can we utilize LLMs to combat misinformation?* On the other hand, the critical challenge is that LLMs can be easily leveraged to generate deceptive misinformation at scale. Then, another important question is: *how to combat LLM-generated misinformation?* In this paper, we first systematically review the history of combating misinformation before the advent of LLMs. Then we illustrate the current efforts and present an outlook for these two fundamental questions respectively. The goal of this survey paper is to facilitate the progress of utilizing LLMs for fighting misinformation and call for interdisciplinary efforts from different stakeholders for combating LLM-generated misinformation<sup>1</sup>.

## 1 Introduction

Misinformation has been a longstanding and serious concern in the contemporary digital age [451]. With the proliferation of social media platforms and online news outlets, the barriers to generating and sharing content have significantly diminished, which also expedites the production and dissemination of various kinds of misinformation (e.g., fake news, rumors) and exaggerates its influence at scale [127, 238, 248, 331, 430, 557, 646, 655]. As the consequence of prevalent misinformation, the public’s belief in truth and authenticity can be under threat. Thus, there is a pressing need to combat misinformation to safeguard information ecosystems and uphold public trust, especially in high-stakes fields such as healthcare [66] and finance [407].

The advent of LLMs [638] (e.g., ChatGPT, GPT-4 [49]) has started to make a transformative impact on the landscape of combating misinformation. In general, LLMs are *a double-edged sword* in the fight against misinformation, indicating that LLMs have brought both emergent *opportunities* and

*challenges*. On the one hand, the profound *world knowledge* and strong *reasoning abilities* of LLMs suggest their potential to revolutionize the conventional paradigms of misinformation *detection*, *intervention* and *attribution*. In addition, LLMs can be augmented with external knowledge, tools, and multimodal information to further enhance their power and can even operate as autonomous agents [563]. On the other hand, the capacities of LLMs to generate human-like content, possibly containing hallucinated information, and follow humans’ instructions [168] indicate that LLMs can be easily utilized to generate misinformation in an *unintentional* or *intentional* way. More seriously, recent research [65] has found that LLM-generated misinformation can be harder to detect for humans and detectors compared to human-written misinformation with the *same semantics*, implying that the misinformation generated by LLMs can have *more deceptive* styles and potentially cause *more harm*.

In this paper, we first provide a comprehensive and systematic review of the history of combating misinformation before the rise of LLMs with a focus on the detection aspect in Section 2. Then we delve into both the opportunities and challenges of combating misinformation in the age of LLMs. As for the opportunities, we will illustrate “*can we utilize LLMs to combat misinformation?*” in Section 3. We will present the motivation for adopting LLMs in the fight against misinformation, the current efforts on utilizing LLMs for combating misinformation, which are mainly around the detection aspect, and an outlook embracing the intervention and attribution aspects. As for the challenges, we will discuss “*how to combat LLM-generated misinformation?*” in Section 4. We will dive into the characterization, emergent threats, and countermeasures of misinformation generated by LLMs. Looking ahead, we also point out the potential real-world devastating risks of LLM-generated misinformation in the near future, which may not be exhibited yet, and the desired interdisciplinary measures. Through this survey paper, we aim to **facilitate the adoption of LLMs in combating misinformation** and call for **collective efforts from stakeholders in different backgrounds to fight misinformation generated by LLMs**.

## 2 History of Combating Misinformation

In this section, we conduct a systematic and comprehensive review of the techniques for detecting online misinformation

<sup>1</sup>More resources on “LLMs Meet Misinformation” are on the website: <https://llm-misinformation.github.io/>

before the emergence of LLMs to provide an overview of the history of combating misinformation in terms of the efforts on detection. Generally, we propose to categorize the detection methods into seven classes based on real-world scenarios: capturing linguistic features, leveraging neural models, exploiting social context, incorporating external knowledge, enhancing generalization ability, minimizing supervision cost, and fusing multilingual and multimodality.

### 2.1 Capturing Linguistic Features

Numerous linguistic features have been studied for differentiating misinformation from true information and can be roughly categorized as *stylistic* features, *complexity* features and *psychological* features [7, 180]. As for *stylistic* features, prior research has found that misleading tweets are usually longer, use a more limited vocabulary, and have more negative sentiment [22, 420]. Also, studies have shown that fake news tends to favor informal, sensational, and affective language style since it aims to attract readers’ attention for a short-term financial or political goal [17, 32, 393]. It is discovered that misleading articles use more swear words, subjective terms, superlatives, and modal adverbs to exaggerate a piece of news [410]. As for *complexity* features, misinformation is likely to be linguistically less complex and more redundant [22] when measured by textual lexical diversity (MTLD) and type-token ratio (TTR) [330]. The typical *psychological* features are based on word counts correlated with different psychological processes and basic sentiment analyses such as Linguistic Inquiry and Word Count (LIWC) dictionaries [485], which are shown to be strongly associated with the possibility of being misleading [323]. Based on the linguistic patterns, multiple detectors are proposed [7, 27, 134, 222, 257, 324, 324, 387]. For example, Mahy-oob et al. proposed to leverage 16 linguistic attributes, which include lexical, grammatical and syntactic features, to identify the nuance between fake and factual news [324].

### 2.2 Leveraging Neural Models

With the development of deep learning in natural language processing, more recent works utilize neural models such as Long Short-Term Memory (LSTM) [177] and Convolutional Neural Network (CNN) [218] for feature extraction and prediction instead of manually extracting linguistic patterns [9, 76, 252, 318, 320, 327, 401, 498, 505, 558, 561]. For example, Chen et al. built an attention-residual network combined with CNN for rumor detection [76]. Vaibhav et al. designed a graph neural network (GNN) based model to capture the sentence-level semantic correlation for fake news detection [498]. Notably, as the burgeoning of pre-trained language models (PLMs), more advanced neural models such as Bidirectional Encoder Representations from Transformers (BERT) [104] are also adopted for misinformation detection [37, 219, 383, 521, 600]. For example, FakeBERT combines BERT and single-layer CNNs with different kernel sizes

and filters as the detector and outperforms conventional machine learning-based models [219].

### 2.3 Exploiting Social Context

Considering social media has been one of the major channels for misinformation production and dissemination, it is essential to incorporate the social context for effectively detecting misinformation and protecting the online information space. Generally, social context can be divided into *social engagements* and *social networks*. The *social engagements* refer to the users’ interactions with content on social media including tweeting, retweeting, commenting, clicking, liking, and disliking. It is found that the user-news interactions are different for fake and authentic news [453]. Thus, a series of works has explored adopting social engagements as useful auxiliary information for detecting misinformation [84, 97, 268, 272, 285, 317, 409, 436, 444, 448, 497, 566, 576, 581, 582, 629]. For example, Shu et al. proposed a sentence-comment co-attention sub-network to jointly model news content and users’ comments for fake news detection [448]. Sheng et al. developed a news environment perception framework to exploit the user-news environment [444]. Another line of works aims to leverage the *social networks*, which encompass multiple concepts such as propagation trajectories, user-user networks, and user-post networks, to enhance the detection performance. Since the structure of social networks can be captured and represented as graphs, a majority of works focus on developing Graph Neural Network (GNN) based models to detect various kinds of misinformation [42, 69, 92, 106, 130, 261, 333, 340, 361, 416, 425, 470, 471, 475, 476, 536, 537, 549, 580, 585, 609]. For example, Wu et al. designed a new graph structure learning approach to leverage the distinctive degree patterns of misinformation on social networks [549]. Jeong et al. developed a hypergraph neural network-based detector to capture the group-level dissemination patterns [470]. Besides graph neural networks, there are also some works modeling social context information with a mixture marked Hawkes model [356], Markov random field [360] or dual-propagation model [246].

### 2.4 Incorporating External Knowledge

There are generally two types of widely used external knowledge embracing *knowledge graphs* and *evidential texts* for assisting misinformation detection. The *knowledge graphs* are usually constructed by domain experts and contain a large number of entities and their relations, which is helpful for checking the veracity of articles [89, 93, 111, 187, 329, 503, 556]. For example, Hu et al. proposed an end-to-end graph neural network to compare the document graph with external knowledge graphs for fake news detection [187]. The *evidential texts* refer to textual facts that can be used for examining the authenticity of articles. Multiple works have investigated evidence-based reasoning strategies for misinformation detection [10, 67, 116, 138, 154, 163, 190, 215,

229, 234, 274, 354, 364, 392, 429, 437, 445, 506, 507, 513, 514, 552, 572, 654]. For example, Jin et al. designed a fine-grained graph-based reasoning framework to incorporate multiple groups of external evidence in the detection process [215].

## 2.5 Enhancing Generalization Ability

In the real world, misinformation can emerge and evolve quickly, indicating that the distribution of misinformation data will likely keep changing. Thus, a line of research works aim to enhance the generalization ability of misinformation detectors under *domain shift* [197, 297, 346, 355, 457, 571, 651] and *temporal shift* [184, 650, 656]. As for *domain shift*, for example, Mosallanezhad et al. built a reinforcement learning-based domain adaptation framework to adapt trained fake news detectors from source domains to target domains [346]. As for temporal shift, one example is that Hu et al. proposed to use the forecasted temporal distribution patterns of news data to guide the misinformation detector [184].

## 2.6 Minimizing Supervision Cost

Another major challenge for misinformation detection in practices is the lack of supervision labels due to the hardness of checking the factuality of articles and the intention to detect misinformation in the early stage of dissemination. Previous works have explored various approaches to address this challenge including data augmentation [169, 452], active learning [118], prompt based learning [198, 286, 551, 617], adversarial contrastive learning [284], transfer learning [251] and meta learning [614]. In particular, multiple works have studied the problem of early misinformation detection [204, 269, 304, 460, 564, 610, 613, 616]. For example, Huang et al. designed a social bot-aware graph neural network to capture bot behaviors for early rumor detection. In addition, there are some other works exploiting the weak supervision signals, which can be weak labels, constraints from heuristic rules, or extrinsic knowledge sources, for misinformation detection [277, 449, 454, 530, 579].

## 2.7 Fusing Multilingual and Multimodality

Recently, it has attracted increasing attention to fuse *multilingual* and *multimodal* information for misinformation detection. As for *multilingual* detection, previous research aim to leverage the high-resource languages to help low-resource languages [86, 95, 99, 107, 194] or build a universal misinformation detector across multiple languages [34, 98, 153, 155, 161, 353, 362, 373, 433, 500]. The *multimodal* detection generally covers various combinations of different modalities including text, images, audio, video, networking and temporal information [1, 14, 59, 75, 77, 83, 85, 128, 189, 225, 291, 293, 294, 348, 394, 399, 438, 474, 474, 478, 479, 481, 527, 540, 547, 562, 602, 620, 633, 640, 641, 648] and has multiple modality fusion strategies including early-fusion, late-fusion and hybrid-fusion [12]. For example, Sun et al. proposed to model the cross-modal and content-knowledge inconsistencies in

a unified framework for multimedia misinformation detection [474]. In particular, combating video misinformation has gained growing interest due to the proliferation of video-sharing platforms such as TikTok and YouTube [48, 83, 395]. One example is that Choi et al. integrated comment, title, and video information with an adversarial learning framework for misinformation detection on YouTube [83].

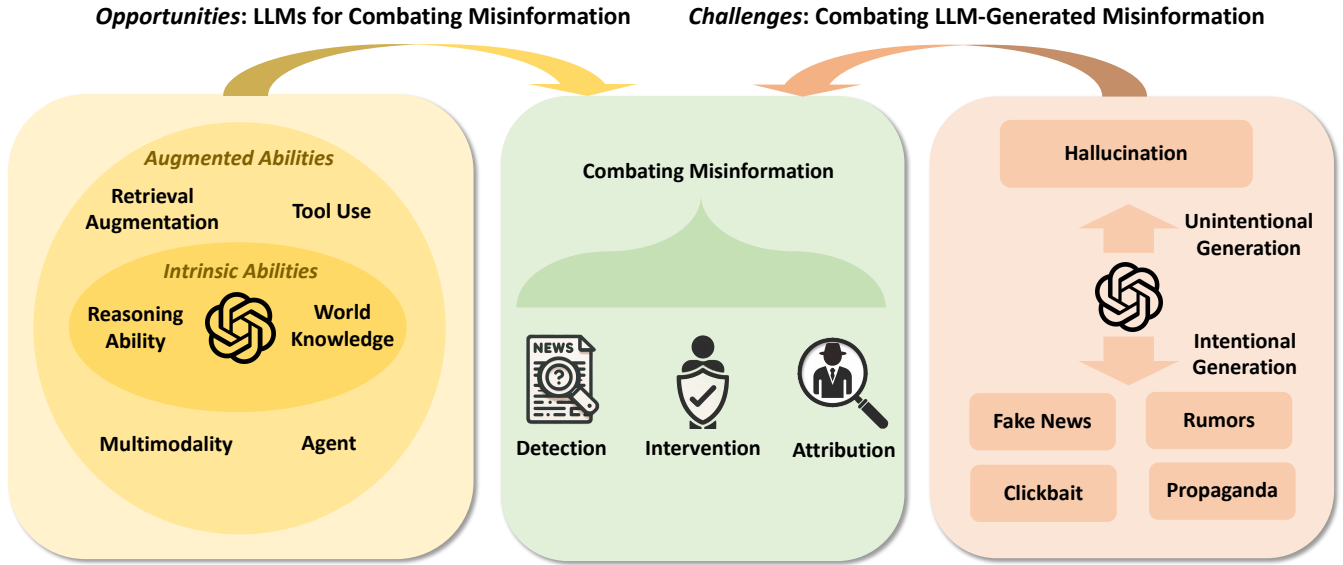
# 3 LLMs for Combating Misinformation

In this section, we aim to illustrate the *opportunities* of combating misinformation in the age of LLMs, *i.e., can we utilize LLMs to combat misinformation?* First, we will introduce the motivation of adopting LLMs in the fight. Then, we will delve into the booming works on leveraging LLMs for misinformation detection. Finally, we will provide an outlook on trustworthy misinformation detection with the assistance of LLMs, utilizing LLMs for misinformation intervention and attribution, and the adoption of multimodal LLMs, LLM agents, and human-LLM collaboration in the future.

## 3.1 Why Adopting LLMs?

Large language models have demonstrated their strong capacities in various tasks such as machine translation [244], summarization [622], and complex question answering [482]. With regard to the realm of combating misinformation, the advent of LLMs has started to revolutionize the previous paradigms of misinformation detection, intervention, and attribution. As shown in Figure 1, we summarize the reasons from three perspectives:

- First, **LLMs contain a significant amount of world knowledge**. Since LLMs are usually pre-trained on a large corpus (e.g., Wikipedia) and have billions of parameters, they can store much more knowledge than a single knowledge graph, which is shown in previous benchmarks [68, 70, 188, 279, 473, 490, 605, 618] and discussed in related surveys [51, 152, 368, 370, 512]. Thus, LLMs have the potential to detect factual errors in misleading texts. One example is shown in Figure 2. Even if “Mercury” and “Aluminum” are medical terminologies, ChatGPT has an accurate understanding of these terms, reflecting that LLMs have a wide range of world knowledge.
- Second, **LLMs have strong reasoning abilities, especially in a zero-shot way**. Previous research has shown that LLMs have powerful capacities in arithmetic reasoning, commonsense reasoning, and symbolic reasoning [87, 191, 400, 568, 608], and can also decompose the problem and reason based on rationales with prompts such as “Let’s think step by step” [231]. Thus, LLMs can potentially reason based on their intrinsic knowledge to determine the authenticity of articles. The example in Figure 2 shows that LLMs such as ChatGPT can reason and explain why a piece of misinformation is misleading. In addition, LLMs’ strong zero-shot reasoning ability also



**Figure 1.** Opportunities and challenges of combating misinformation in the age of LLMs.

largely solves the challenges of distribution shifts and lack of supervision labels in the real world.

- Third, **LLMs can be augmented with external knowledge, tools, and multimodal information, and can even operate as autonomous agents.** One major limitation of LLMs is that they can potentially generate hallucinations, which refer to the LLM-generated texts containing nonfactual information. One of the main reasons for hallucinations is that LLMs cannot get access to up-to-date information and may have insufficient knowledge in specialized domains such as healthcare [210, 414, 632]. Recent research has shown that LLMs’ hallucinations can be mitigated with the augmentation of retrieved external knowledge [296, 298, 352, 439, 625] or tools (e.g., search engines such as Google) to get access to up-to-date information [131, 143, 200, 398, 402, 403, 611]. Furthermore, LLMs can be tuned to reason based on multimodal information including images, code, tables, audio, and graphs [588, 637], which indicates LLMs can also be applicable to combating multimodal misinformation. LLMs have also been shown to have the capacity to serve as autonomous agents in various tasks [21, 301, 309, 519, 563, 573], which has great potential to be used for autonomizing the process of fact-checking and misinformation detection.

### 3.2 LLMs for Misinformation Detection

Recently, it has already witnessed increasing efforts exploring how to utilize LLMs for misinformation detection. Initially, some works have investigated directly prompting GPT-3<sup>2</sup> [50, 271], InstructGPT [369], ChatGPT-3.5<sup>3</sup> [33, 53, 178,

202, 213, 232, 276, 516, 627, 630] and GPT-4<sup>4</sup> [65, 384, 405] for misinformation detection. For example, Pan et al. [369] presented a program-guided fact-checking framework that leverages the in-context learning ability of LLMs to generate reasoning programs to guide veracity verification. Chen et al. [65] have studied ChatGPT-3.5 and GPT-4 with the standard prompting (“No CoT”) strategy and zero-shot chain-of-thought (“CoT”) prompting strategy for both human-written misinformation and LLM-generated misinformation. The extensive experiments show that the “CoT” strategy mostly outperforms the “No CoT” strategy. Also, a few recent works have started to leverage LLMs for detecting multimodal misinformation. One example is that Wu et al. used GPT-3.5 as the feature extractor to detect out-of-context images [546]. Besides directly prompting LLMs, Pavlyshenko et al. [381] adopted the parameter-efficient fine-tuning LoRA [185] on an open-sourced LLM Llama 2 [488] for multiple tasks including fact-checking and fake news detection.

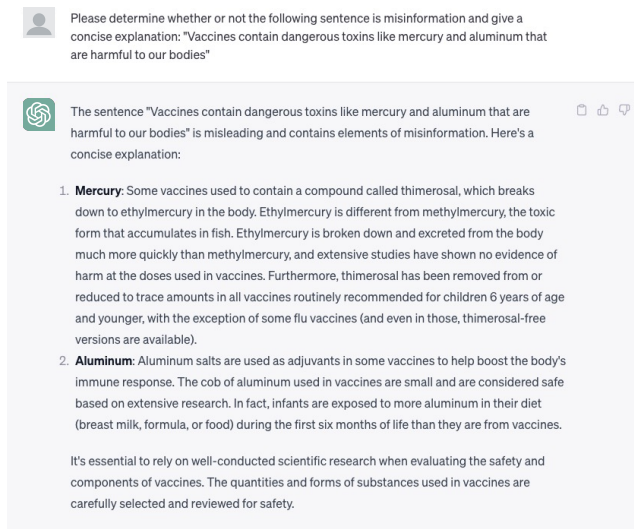
Since the knowledge contained in LLMs may not be up-to-date or sufficient in detecting factual errors, some works have explored augmenting LLMs with external knowledge [82] or tools [81] for misinformation detection. Specifically, Cheung et al. combined the retrieved knowledge from a search engine and the reasoning ability of Llama to predict the veracity of claims [82]. Chern et al. proposed a fact-checking framework integrated with multiple tools (e.g., Google Search, Google Scholar, code interpreters, Python) to detect the factual errors of texts generated by LLMs [81]. In addition, some works studied utilizing LLMs to assist conventional supervisedly trained detectors via generating weak labels [256], rationales [183] or instances [6, 193, 456]. For example, Leite

<sup>2</sup>gpt-3: <https://platform.openai.com/docs/models/gpt-3>

<sup>3</sup>gpt-3.5: <https://platform.openai.com/docs/models/gpt-3-5>

<sup>4</sup>gpt-4: <https://platform.openai.com/docs/models/gpt-4>





**Figure 2.** An example of leveraging ChatGPT to detect misinformation and give explanations.

et al. employed a weakly-supervised learning framework Snorkel [411] to leverage LLM-generated supervision signals for training misinformation detectors [256].

### 3.3 Outlook

In this subsection, we provide an outlook on combating misinformation in the age of LLMs. First, we can further harness multilingual and multimodal LLMs to build effective and trustworthy detectors. Second, although the existing works mainly focus on the detection of misinformation, LLMs have great potential to be adopted in misinformation intervention and attribution. In addition, we will discuss the application of human-LLM collaboration in combating misinformation.

**3.3.1 Trustworthy Misinformation Detection.** Though previous misinformation detectors have achieved relatively high performance, it is under exploration on how to ensure trustworthiness in the detection process including robustness, explainability, fairness, privacy, and transparency, which is essential for gaining the public trust [79, 292, 295, 487]. Some previous works have explored the robustness [249, 311, 316, 461, 515] and explainability [126, 226, 267, 313, 448, 576, 589] of misinformation detectors. However, all these works are based on conventional supervisedly trained detectors, the emergence of LLMs has brought new opportunities for building trustworthy detectors. For example, as shown in Figure 2, LLMs such as ChatGPT can generate fluent natural language-based explanations for the given misinformation while predicting the authenticity, which is more human-friendly than previous extraction-based explanation methods [448]. The other aspects of trustworthiness for LLM-based detectors are still under study.

### 3.3.2 Harnessing Multilingual and Multimodal LLMs.

It has been demonstrated that LLMs can be naturally extended to multilingual languages [28, 78, 136, 484, 538, 539, 545] and multimodalities [72, 125, 259, 260, 308, 517, 525, 588, 596, 624, 628]. First, multilingual LLMs have shown strong generalization ability across different languages including many low-resource ones. For example, one LLM named Phoenix [78] can generalize to both Latin (e.g., Deutsch) and non-Latin languages (e.g., Arabic). Thus, multilingual LLMs can largely alleviate the low-resource challenges in cross-lingual misinformation detection. Second, recent studies have demonstrated the impressive multi-sensory skills of multimodal LLMs [596]. In particular, GPT-4V [588] has manifested surprising capacities of visual-language understanding and reasoning, indicating the profound potential in combating multimodal misinformation in the real world.

### 3.3.3 LLMs for Misinformation Intervention.

Different from *misinformation detection* methods that mainly focus on checking the veracity of given texts, *misinformation intervention* approaches go beyond the pure algorithmic solutions and aim to exert a direct influence on users [4, 31, 165, 427, 428], which is also a critical component of the lifecycle of combating misinformation. Generally, there are two lines of intervention measures. The most standard intervention measures follow the pipeline of fact-checking and debunking after humans are already exposed to the misinformation [62, 216, 382, 382, 509, 510, 603]. The potential usage of LLMs is to improve the convincingness and persuasive power of the debunking responses. For example, He et al. [166] proposed to combine reinforcement learning and GPT-2 to generate polite and factual counter-misinformation responses. However, one drawback of these post-hoc intervention methods is that they may cause a psychological “backfiring effect”, suggesting that humans end up believing more in the original misinformation [103, 258, 480]. Thus, another line of intervention methods aims to leverage inoculation theories to immunize the public against misinformation [499]. Karinshak et al. pointed out the potential of employing LLMs to generate persuasive anti-misinformation messages (e.g., pro-vaccination messages) in advance to enhance the public’s immunity against misinformation [223].

### 3.3.4 LLMs for Misinformation Attribution.

Misinformation attribution refers to the task of identifying the author or source of given misinformation, based on the assumption that the texts written by different authors are likely to have distinct stylometric features and these features will be preserved in different texts for the same author [493, 494]. Misinformation attribution plays a vital role in combating misinformation because it can be leveraged to trace the origin of propaganda or conspiracy theories and hold the publishers accountable. Although there are still no works adopting LLMs in misinformation attribution, LLMs have already

exhibited great power in identifying [378] and manipulating [314, 377, 415, 421] stylometric features, indicating the promise for tracing the authorship of misinformation. For example, Patel et al. [378] performed stylometric analysis on a large number of texts via prompting GPT-3 and created a human-interpretable stylometry dataset, which shows that LLMs have a deep understanding of the stylometric features.

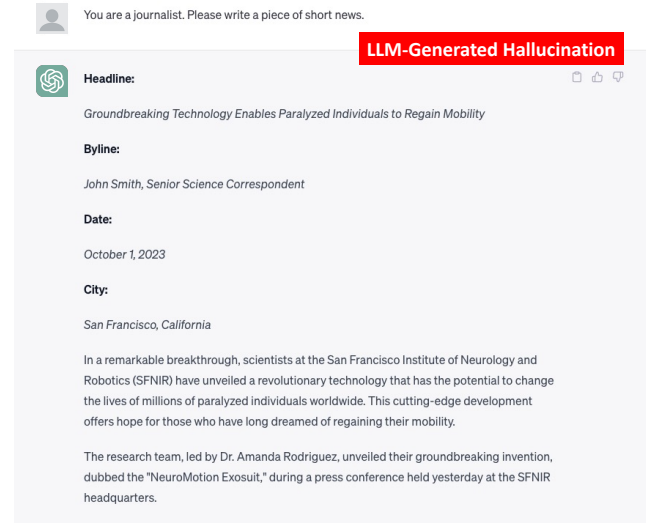
**3.3.5 Human-LLM Collaboration.** The research in the realm of human-AI collaboration and teaming aims to leverage the strengths of both humans and AI [18, 19, 181]. First, human guidance helps steer the development of AI to maximize AI’s benefits to humans and ensure AI will not cause unintended harm, especially for minority groups. Second, AI can boost humans’ analytic and decision-making abilities by providing useful auxiliary information. There are already some works studying the adoption of human-AI collaboration in combating misinformation [233, 336, 359, 462, 495]. For example, Mendes et al. proposed a human-in-the-loop evaluation framework for early detection of COVID-19 misinformation on social media, which combines both modern NLP methods and experts’ involvement [336]. In the age of LLMs, we call for more research to leverage the best of both LLMs and humans in fighting misinformation.

## 4 Combating LLM-Generated Misinformation

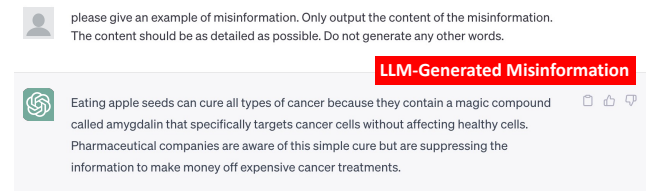
In this section, we will delve into the emerging *challenges* in the age of LLMs. *i.e.*, **how to combat LLM-generated misinformation?** First, we will provide a characterization of misinformation generated by LLMs, which has attracted increasing attention in recent works [26, 65, 115, 141, 142, 160, 162, 202, 213, 228, 372, 463, 468, 477, 535, 550, 644]. Then, we will illustrate the new threats brought by LLM-generated misinformation in different fields, and the countermeasures against LLM-generated misinformation including alleviating hallucination, improving LLMs’ safety and detecting LLM-generated misinformation. Finally, looking ahead, we anticipate the misinformation generated by LLMs and other large generative AI models will cause more devastating real-world impacts, which may not appear yet. Thus, we will discuss the potential risks in the near future and the desired measures.

### 4.1 Characterization

In general, LLM-generated misinformation can be divided into *unintentional generation* and *intentional generation* by different intents [65]. As shown in Figure 1, the misinformation generated via *unintentional generation* methods mainly refers to hallucinations, *i.e.*, the nonfactual texts generated by LLMs. Since hallucinations can occur in any generation process of LLMs due to the intrinsic properties of auto-regressive generation and lack of up-to-date information [210, 414, 594, 632], users without malicious intents may also generate non-factual content when prompting LLMs. An example is shown



**Figure 3.** An example of *unintentionally* prompting ChatGPT to generate misinformation (*i.e.*, hallucination).



**Figure 4.** An example of *intentionally* prompting ChatGPT to generate Misinformation.

in Figure 3. When users adopt prompts such as “please write a piece of short news”, LLMs (*e.g.*, ChatGPT) will probably generate content containing hallucinated information, in particular the fine-grained information such as dates, names, addresses, numbers, and quotes, even if the main message may seem to be correct (*e.g.*, the “Byline”, “Date” and “City” in Figure 3 are fabricated). The *intentional generation* methods suggest that malicious users can knowingly prompt LLMs to generate various kinds of misinformation including fake news, rumors, conspiracy theories, clickbait, misleading claims, or propaganda. One example is shown in Figure 4. When the users use prompts such as “please give an example of misinformation ...”, LLMs (*e.g.*, ChatGPT) potentially will generate a piece of misinformation such as “Eating apple seeds can cure all types of cancer because they contain a magic compound called amygdalin ...”, though LLMs can also possibly reply with “As an AI language model, I cannot provide misinformation” owing to the intrinsic safety guard mechanisms of LLMs. Notably, recent research [65] has found that the misinformation generated by LLMs (*e.g.*, ChatGPT) can be **harder to detect** for *humans* and *detectors* compared with human-written misinformation with the **same semantics**, indicating that LLM-generated misinformation can have

*more deceptive styles* and potentially cause more harm. Another work [463] also shows that GPT-3 can generate both accurate information that is easier to understand and misinformation that is more compelling.

## 4.2 Emerging Threats

LLM-generated misinformation has already posed serious threats in the real world [36, 38, 122, 142, 335, 458, 542, 543]. In this subsection, we will discuss the immediate threats of the LLM-generated misinformation on a variety of fields including journalism, healthcare, finance, and politics considering its characteristics of *deceptiveness* and *easy production*.

**4.2.1 Journalism.** Journalism may be one of the fields that LLM-generated misinformation has the most substantial impact on. For example, in April 2023, NewsGuard identified 49 LLM-powered news websites in 7 languages including English, Chinese, Czech, French, Portuguese, Tagalog, and Thai [358]. These websites can possibly produce hundreds of clickbait articles a day to optimize the advertisement revenue, which causes vast amounts of pollution to online information ecosystems. Since LLM-generated misinformation can have more deceptive styles than human-written misinformation with the same semantics [65], it is challenging for readers, fact-checkers, and detection algorithms to effectively discern truth from the misleading information generated by LLMs. In the long run, as the line between human-written news and LLM-generated news blurs, the public trust in legitimate news sources could be undermined and the journalistic ethos – centered on accuracy, accountability, and transparency – might be put to the test. Thus, it is imperative for news outlets to guarantee authenticity and uphold public trust.

**4.2.2 Healthcare.** Recent works have pointed out the rise of adoption of LLMs in healthcare applications [167, 221, 266, 299, 357, 423, 426], however, they can also inadvertently be a tool for the generation and propagation of health misinformation [96]. For example, it is found that LLMs such as GPT-3 can be used to generate totally fabricated health articles that appear remarkably authentic [325]. Compared with the informatics driven by human-written misinformation [13, 56, 66, 245, 380, 388], it can be even tougher to combat **LLM-driven informatics** for the following reasons. *First*, it is hard for unsuspecting users, who may lack the nuanced understanding of clinical context and medical research, to distinguish LLM-generated hallucinated health content from authentic medical information. If they rely on LLMs to seek health advice, it may lead to potential misinterpretations and adverse health outcomes. *Second*, malicious actors can manipulate LLMs to craft plausible-sounding yet erroneous medical content, promoting alternative healthcare treatments or disproven theories for profit. This will not only undermine the credibility of genuine health information but also pose significant risks to public health.

**4.2.3 Finance.** Previous research has shown that human-written financial misinformation can cause various detrimental consequences such as disrupting markets, misleading investors, and amplifying economic instability [407]. In the age of LLMs, the financial sector faces an even more escalating threat from LLM-generated misinformation, because bad actors, who are potentially motivated by profit, sabotage, or other malicious intents, can easily leverage LLMs to spread disinformation campaigns, create counterfeit financial statements, or even impersonate legitimate financial analysis. Furthermore, considering the prevalence of high-frequency trading and algorithm-driven investment decisions, even short-lived misinformation may trigger automated trades misled by the fabricated content. For example, it is reported that the stock price of an artificial intelligence company iFlytek has a deep drop due to a piece of chatbot-generated misinformation [434]. Thus, stakeholders in the financial industry should increase their awareness of the potential threats of LLM-generated misleading content.

**4.2.4 Politics.** Misinformation has a longstanding grave impact on the political spheres [2, 8, 135, 144, 159, 217, 328, 345, 374, 390, 391, 466]. The advent of LLMs can potentially usher in a new age of misinformation and disinformation in the realm of politics. The reasons can be summarized as the following two points. The first threat of LLM-generated misinformation is **distorting democracy**. LLMs can be easily weaponized to generate deceptive narratives about candidates, policies, or events at scale. When people are exposed to such content, their perception of election candidates might be altered, leading them to vote differently. More seriously, the flooding of LLM-generated misinformation can possibly weaken the citizens’ trust in the whole democratic process and eventually erode the foundations of democratic systems. The second threat is **amplifying polarization**. Bad actors may leverage LLMs to craft personalized misinformation tailored to individual biases and beliefs, which may resonate with specific audiences and increase the likelihood of spreading among targeted communities. This can result in exacerbating echo chambers and confirmation biases, driving wedges between different groups and making consensus-building even harder in the political spheres.

## 4.3 Countermeasures

In this subsection, we will discuss four major countermeasures against LLM-generated misinformation including alleviating hallucination of LLMs, improving the safety of LLMs, detecting LLM-generated misinformation, and public education, through which we hope to inspire more future works on combating LLM-generated misinformation.

**4.3.1 Alleviating Hallucination of LLMs.** Hallucination is the main source of unintentional LLM-generated misinformation. Recently, increasing works start to design approaches for evaluating [80, 81, 108, 210, 265, 287, 289, 349,

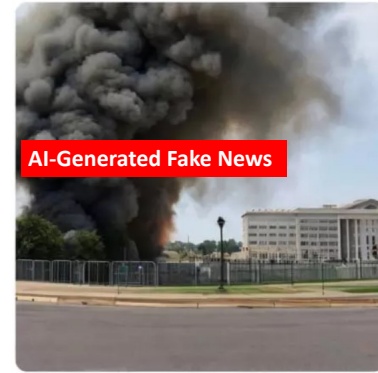


413, 414, 464, 592, 594, 597, 623] or mitigating hallucination [5, 112, 114, 119, 147, 211, 255, 270, 290, 315, 326, 351, 626, 635]. In general, there are two lines of works on hallucination mitigation. In the *training* stage, previous research has explored training data curation or knowledge grounding methods to incorporate more knowledge [186, 370, 607, 647]. In the *inference* stage, recent works have investigated methods including confidence estimation (or uncertainty estimation) [201, 501, 567], knowledge retrieval [120, 214, 264, 306, 338, 508, 526, 601], enabling citations [132, 192], model editing [275, 337, 424, 593], multi-agent collaboration [90, 110], prompting [105, 544] and decoding strategy design [88, 253].

**4.3.2 Improving Safety of LLMs.** The safety guard of LLMs aims to prevent malicious users exploiting LLMs to generate harmful content including various types of misinformation, which has been emphasized in a variety of survey papers [15, 44, 52, 63, 64, 73, 91, 101, 102, 113, 117, 151, 171, 196, 224, 237, 281, 305, 347, 435, 446, 522, 529, 548, 577]. A line of works has evaluated or benchmarked the safety of various LLMs [203, 205, 288, 386, 412, 511, 523, 524, 569, 595, 621, 636, 649, 653]. Generally, the safety of LLMs can also be strengthened in both *training* and *inference* stage. In the *training* stage, previous works focus on designing alignment training approaches such as reinforcement learning from human feedback (RLHF) to align LLMs with humans’ values [29, 30, 43, 121, 209, 242, 365, 441, 472, 492, 531, 591, 642, 643]. In the *inference* stage, existing research has studied red teaming methods to find LLMs’ flaws [58, 129, 250, 332, 334, 385, 447, 465, 598, 604, 612, 652], prompt injection or jailbreak approaches to probe LLMs’ safety risks [100, 145, 199, 220, 247, 263, 300, 302, 303, 396, 404, 408, 443, 574, 590], and defense methods for the evolving jailbreaks [170, 172, 236, 417, 541].

**4.3.3 Detecting LLM-Generated Misinformation.** Misinformation detection is an important measure for platforms to prevent its dissemination, which has been discussed in related surveys [16, 41, 45, 54, 66, 148, 206, 239, 363, 375, 440, 496, 619, 631]. Previously, there are a large number of works on detecting human-written misinformation including fake news [23, 24, 164, 208, 280, 283, 459, 518, 553, 634], rumor [133, 273, 319, 379], clickbait [74], cherry-picking [25], and propaganda [94, 322, 327]. Recently, more research focuses on machine-generated misinformation or neural misinformation, suggesting that it is generated by neural models, such as [3, 7, 39, 109, 128, 162, 249, 406, 450, 615] and its detection methods [40, 367, 432, 463, 467, 481]. In the age of LLMs, there start to be some initial works exploring LLM-generated misinformation detection [65, 81, 115, 142, 160, 213, 372, 532, 550, 644], but more research is strongly desired. It is worth noting that detecting LLM-generated misinformation holds a close connection with the techniques in detecting LLM-generated texts, which can be directly adopted in detecting LLM-generated misinformation or take effect via notifying the readers of the potential inauthenticity. The problem of

Reports of an explosion near the Pentagon in Washington DC



12:03 am · 23/5/2023 from Earth · 38.2K Views

135 Retweets 50 Quotes 354 Likes

**Figure 5.** A real-world example of AI-Generated Multimodal Misinformation.

detecting LLM-generated texts [60, 137, 150, 174, 175, 240, 241, 243, 278, 282, 307, 321, 341, 343, 344, 350, 422, 469, 483, 486, 495, 502, 504, 528, 534, 554, 555, 570, 584, 586, 606] as well as the watermarking techniques [182, 230, 235, 254, 520, 575, 583, 599, 639] has attracted increasing attention.

**4.3.4 Public Education.** The goal of public education is two-fold. First, the general public should be educated about the capacities and limitations of LLMs, which can include the understanding that while LLMs can produce coherent and plausible-sounding texts, the LLM-generated content may contain nonfactual information. Thus, the public education can potentially reduce the risk of normal people abusing LLMs and generating profound hallucinated information unintentionally. Second, it is imperative to enhance the public’s digital literacy and immunity against LLM-generated misinformation. For example, the characteristics of LLM-generated misinformation and the identification approaches should be taught in different communities, especially the minority groups who have been found to be more susceptible to misinformation [207, 227, 310, 366, 371].

## 4.4 Looking Ahead

In this subsection, we will discuss the potential risks of misinformation generated by LLMs as well as other large generative AI models in the near future, which may not explicitly be exhibited yet, including AI-generated multimodal misinformation, autonomous misinformation agents, cognitive security and AI-manipulation, as well as the needed interdisciplinary countering measures.

**4.4.1 AI-Generated Multimodal Misinformation.** With the development of generative AI, we have witnessed an exponential increase of various tools to create content in

multimodalities, which include not only texts but also audio, images, and video. For example, users can create high-resolution images with close-sourced (e.g., Midjourney [339]) or open-sourced (e.g., Stable Diffusion [419]) text-to-image generation tools. Also, multimodal LLMs (e.g., GPT-4V [588]) have demonstrated surprisingly strong capacities for visual understanding and image-to-text generation. In reality, malicious actors can easily combine these tools to craft hyper-realistic yet entirely fabricated multimodal misinformation, which may bring more challenges for normal people and even digital experts. A real-world example of AI-generated multimodal fake news is shown in Figure 5, which contains both a piece of misleading text “Reports of an explosion near the Pentagon in Washington DC” and a synthetic fake image. We can also see the extent of potential impact from the number of “views” and “likes”.

**4.4.2 Autonomous Misinformation Agents.** Recent advances in LLM agents have shown that LLMs can finish a wide range of complex tasks automatically which require multiple human-level abilities including planning, reasoning, executing, reflecting, and collaborating [71, 179, 309, 519, 533, 559, 563, 645]. In the future, we can envision a society where humans and agents live together [262]. However, it is also shown that the current safety guard of LLMs can be easily broken via fine-tuning [397, 455, 587]. Thus, the bad actors can possibly create malicious autonomous misinformation agents and deploy them in online information ecosystems. The potential danger is that these misinformation agents may operate without the need for humans’ detailed instructions, tirelessly generate vast amounts of misleading content, adapt to conversational contexts in real-time, and adjust their messages to cater to specific targeted audiences, which will make a devastating impact on public trust and online safety. It is worth noting that some recent works also discuss the risks of LLM-powered bots [123, 578] and agentic systems [61]. To ensure that humans and agents live in harmony in the future, more multidisciplinary efforts are desired.

**4.4.3 Cognitive Security and AI-Manipulation.** The ultimate goal of AI technologies including LLMs should be to maximize the benefits for humans. However, in the future, LLM-generated misinformation could be weaponized to serve as an emerging type of AI-powered **cognitive attacks**, which can be defined as the cyber-physical-human processes that manipulate humans’ behaviors for malicious purposes by exploiting their cognitive vulnerabilities [195], which pose serious concerns to humans’ **cognitive security** [149]. Recent evidence has shown that LLMs can be leveraged to infer the cognitive properties (e.g., personalities) of humans from social media posts [389]. It is possible that bad actors or LLM-powered autonomous misinformation agents may exploit humans’ cognitive vulnerabilities to maximize the impact, which is especially concerning for minority communities. Furthermore, LLM-generated

misinformation can also be regarded as a new kind of **AI-manipulation** [55, 57, 173, 376] or **social media manipulation** [11, 565]. It is under-explored how to protect humans against the negative impact of LLM-generated misinformation from a cognitive perspective.

**4.4.4 Interdisciplinary Countering Efforts.** In the long run, combating LLM-generated misinformation needs efforts from different disciplines including technology, sociology, psychology, education, and policymaking. From the **technology** perspective, first, the factuality and safety aspects of LLMs should be further strengthened. Second, more effective detection methods for LLM-generated misinformation or texts are strongly needed. From the **sociology** perspective, understanding the patterns of the dissemination of LLM-generated misinformation or the behavior of LLM-powered misinformation agents can help prevent the spread. From the **psychology** perspective, recognizing the cognitive weaknesses that make individuals susceptible to misinformation, which may be exploited by bad actors or LLM agents, can lead to more effective intervention measures. From the **education** perspective, courses on digital literacy and critical thinking can enhance the public’ discernment skills on LLM-generated misinformation. From the **policymaking** perspective, it is pressing to enact regulations to mandate transparency and accountability in the development and deployment of both close-sourced (e.g., ChatGPT, GPT-4) or open-sourced (e.g., Llama2 [488], Mistral [212]) LLMs. It is worth noting that the regulation of LLMs is an important component of the overall picture of AI regulation, which is also discussed in recent works [20, 35, 46, 47, 124, 139, 140, 146, 156–158, 176, 312, 342, 418, 431, 442, 489, 491, 560]. In addition, we also need to involve the **general public** in the fight against LLM-generated misinformation and foster constructive discussions on the implications of LLM-generated misinformation on free speech, privacy, and other fundamental rights. By harnessing the efforts of multiple disciplines and different stakeholders, we can form a multi-pronged defense framework to combat LLM-generated misinformation and safeguard information ecosystems.

## 5 Conclusion

The advent of LLMs can potentially usher in a new era of combating misinformation, indicating both emergent opportunities and challenges. This survey paper first provides a systematic review of the history of combating misinformation before the rise of LLMs. Then, we dive into an in-depth discussion on the existing efforts and future outlook around two fundamental questions on combating misinformation in the age of LLMs: *can we utilize LLMs to combat misinformation* and *how to combat LLM-generated misinformation*. Overall, LLMs have great potential to be adopted in the fight against misinformation, and more efforts are needed to minimize the risks of LLM-generated misinformation.

## References

- [1] Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. 2022. Open-Domain, Content-based, Multi-modal Fact-checking of Out-of-Context Images via Online Resources. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 14920–14929. <https://doi.org/10.1109/CVPR52688.2022.01452>
- [2] A. Abilov, Yiqing Hua, Hana Matatov, Ofra Amir, and Mor Naaman. 2021. VoterFraud2020: a Multi-modal Dataset of Election Fraud Claims on Twitter. *International Conference on Web and Social Media* (2021). <https://doi.org/10.1609/icwsm.v15i1.18113>
- [3] David Ifeoluwa Adelani, Haotian Mai, Fuming Fang, Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. 2019. Generating Sentiment-Preserving Fake Online Reviews Using Neural Language Models and Their Human- and Machine-based Detection. *arXiv preprint arXiv: 1907.09177* (2019).
- [4] Zhila Aghajari, Eric P. S. Baumer, and Dominic DiFranzo. 2023. Reviewing Interventions to Address Misinformation: The Need to Expand Our Vision Beyond an Individualistic Focus. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 87 (2023), 34 pages. <https://doi.org/10.1145/3579520>
- [5] Ayush Agrawal, Lester Mackey, and Adam Tauman Kalai. 2023. Do Language Models Know When They’re Hallucinating References? *arXiv preprint arXiv: 2305.18248* (2023).
- [6] Emil Ahlback and Max Dougly. 2023. Can Large Language Models Enhance Fake News Detection?: Improving Fake News Detection With Data Augmentation.
- [7] Ankit Aich, Souvik Bhattacharya, and Natalie Parde. 2022. Demystifying Neural Fake News via Linguistic Feature-Based Interpretation. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 6586–6599. <https://aclanthology.org/2022.coling-1.573>
- [8] Rachith Aiyappa, Matthew R. DeVerna, Manita Pote, Bao Tran Truong, Wanying Zhao, David Axelrod, Aria Pessianzadeh, Zoher Kachwala, Munjung Kim, Ozgur Can Seckin, Minsuk Kim, Sunny Gandhi, Amrutha Manikonda, Francesco Pierri, Filippo Menczer, and Kai-Cheng Yang. 2023. A Multi-Platform Collection of Social Media Posts about the 2022 U.S. Midterm Elections. *arXiv preprint arXiv: 2301.06287* (2023).
- [9] Oluwaseun Ajao, Deepayan Bhowmik, and Shahrzad Zargari. 2019. Sentiment Aware Fake News Detection on Online Social Networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*. IEEE, 2507–2511. <https://doi.org/10.1109/ICASSP.2019.8683170>
- [10] Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. 2022. PubHealthTab: A Public Health Table-based Dataset for Evidence-based Fact Checking. In *Findings of the Association for Computational Linguistics: NAACL 2022*. Association for Computational Linguistics, Seattle, United States, 1–16. <https://doi.org/10.18653/v1/2022.findings-naacl.1>
- [11] Mohammad Majid Akhtar, Rahat Masood, Muhammad Ikram, and Salil S. Kanhere. 2023. False Information, Bots and Malicious Campaigns: Demystifying Elements of Social Media Manipulations. *arXiv preprint arXiv: 2308.12497* (2023).
- [12] Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022. A Survey on Multimodal Disinformation Detection. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 6625–6643. <https://aclanthology.org/2022.coling-1.576>
- [13] Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, Abdulaziz Al-Homaid, Wajdi Zaghouani, Tommaso Caselli, Gijs Danoe, Friso Stolk, Britt Bruntink, and Preslav Nakov. 2021. Fighting the COVID-19 Infodemic: Modeling the Perspective of Journalists, Fact-Checkers, Social Media Platforms, Policy Makers, and the Society. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 611–649. <https://doi.org/10.18653/v1/2021.findings-emnlp.56>
- [14] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. *DEEPMIND* (2022).
- [15] Joshua Albrecht, Ellie Kitanidis, and Abraham J. Fetterman. 2022. Despite “super-human” performance, current LLMs are unsuited for decisions about ethics and safety. *arXiv preprint arXiv: 2212.06295* (2022).
- [16] Ihsan Ali, Mohamad Nizam Bin Ayub, Palaiahnakote Shivakumara, and Nurul Fazmidar Binti Mohd Noor. 2022. Fake News Detection Techniques on Social Media: A Survey. *Wireless Communications and Mobile Computing* 2022 (2022).
- [17] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives* 31, 2 (2017), 211–236.
- [18] Saleema Amershi. 2011. Designing for effective end-user interaction with machine learning. In *Proceedings of the 24th annual ACM symposium adjunct on User interface software and technology*. 47–50.
- [19] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *Ai Magazine* 35, 4 (2014), 105–120.
- [20] Markus Anderljung, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O’Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, Ben Chang, Tatum Collins, Tim Fisk, Gillian Hadfield, Alan Hayes, Lewis Ho, Sara Hooker, Eric Horvitz, Noam Kolt, Jonas Schuett, Yonadav Shavit, Divya Siddarth, Robert Trager, and Kevin Wolf. 2023. Frontier AI Regulation: Managing Emerging Risks to Public Safety. *arXiv preprint arXiv: 2307.03718* (2023).
- [21] Jacob Andreas. 2022. Language Models as Agent Models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 5769–5779. <https://aclanthology.org/2022.findings-emnlp.423>
- [22] Dimosthenis Antypas, Jose Camacho-Collados, Alun Preece, and David Rogers. 2021. COVID-19 and Misinformation: A Large-Scale Lexical Analysis on Twitter. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*. Association for Computational Linguistics, Online, 119–126. <https://doi.org/10.18653/v1/2021.acl-srw.13>
- [23] Antonio A Arechar, Jennifer Allen, Adam J Berinsky, Rocky Cole, Ziv Epstein, Kiran Garimella, Andrew Gully, Jackson G Lu, Robert M Ross, Michael N Stagnaro, et al. 2023. Understanding and combatting misinformation across 16 countries on six continents. *Nature human behaviour* (2023), 1–12.
- [24] Arnav Arora, Preslav Nakov, Momchil Hardalov, Sheikh Muhammad Sarwar, Vibha Nayak, Yoan Dinkov, Dimitrina Zlatkova, Kyle Dent, Ameysa Bhatawdekar, Guillaume Bouchard, and Isabelle Augenstein. 2021. Detecting Harmful Content On Online Platforms: What Platforms Need Vs. Where Research Efforts Go. *Comput. Surveys* (2021). <https://doi.org/10.1145/3603399>

- [25] Abolfazl Asudeh, Hosagrahar Visvesvaraya Jagadish, You Wu, and Cong Yu. 2020. On detecting cherry-picked trendlines. *Proceedings of the VLDB Endowment* 13, 6 (2020), 939–952.
- [26] Navid Ayoobi, Sadat Shahriar, and Arjun Mukherjee. 2023. The Looming Threat of Fake and LLM-generated LinkedIn Profiles: Challenges and Opportunities for Detection and Prevention. In *Proceedings of the 34th ACM Conference on Hypertext and Social Media*. 1–10.
- [27] Lucas Azevedo, Mathieu d’Aquin, Brian Davis, and Manel Zarrouk. 2021. LUX (Linguistic aspects Under eXamination): Discourse Analysis for Automatic Fake News Classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, Online, 41–56. <https://doi.org/10.18653/v1/2021.findings-acl.4>
- [28] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen Technical Report. *arXiv preprint arXiv: 2309.16609* (2023).
- [29] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv preprint arXiv: Arxiv-2204.05862* (2022).
- [30] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, John Kernion, Andy Jones, A. Chen, Anna Goldie, Azalia Mirhoseini, C. McKinnon, Carol Chen, Catherine Olsson, C. Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, E. Perez, Jamie Kerr, J. Mueller, Jeff Ladish, J. Landau, Kamal Ndousse, Kamile Lukošiušė, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemí Mercado, Nova DasSarma, R. Lasenby, Robin Larson, Sam Ringer, Scott Johnston, S. Kravec, Sheer El Showk, Stanislaw Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, T. Henighan, Tristan Hume, Sam Bowman, Zac Hatfield-Dodds, Benjamin Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom B. Brown, and Jared Kaplan. 2022. Constitutional AI: Harmlessness from AI Feedback. *ARXIV.ORG* (2022). <https://doi.org/10.48550/arXiv.2212.08073>
- [31] Joseph B Bak-Coleman, Ian Kennedy, Morgan Wack, Andrew Beers, Joseph S Schafer, Emma S Spiro, Kate Starbird, and Jevin D West. 2022. Combining interventions to reduce the spread of viral misinformation. *Nature Human Behaviour* 6, 10 (2022), 1372–1380.
- [32] Vian Bakir and Andrew McStay. 2018. Fake news and the economy of emotions: Problems, causes, solutions. *Digital journalism* 6, 2 (2018), 154–175.
- [33] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. *arXiv preprint arXiv: Arxiv-2302.04023* (2023).
- [34] Giorgio Barnabò, Federico Siciliano, Carlos Castillo, Stefano Leonardi, Preslav Nakov, Giovanni Da San Martino, and Fabrizio Silvestri. 2022. FbMultiLingMisinfo: Challenging Large-Scale Multilingual Benchmark for Misinformation Detection. In *2022 International Joint Conference on Neural Networks (IJCNN)*. 1–8. <https://doi.org/10.1109/IJCNN55064.2022.9892739>
- [35] Anthony M. Barrett, Dan Hendrycks, Jessica Newman, and Brandie Nonnecke. 2022. Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks. *arXiv preprint arXiv: 2206.08966* (2022).
- [36] Clark Barrett, Brad Boyd, Ellie Burzstein, Nicholas Carlini, Brad Chen, Jihye Choi, Amrita Roy Chowdhury, Mihai Christodorescu, Anupam Datta, Soheil Feizi, Kathleen Fisher, Tatsunori Hashimoto, Dan Hendrycks, Somesh Jha, Daniel Kang, Florian Kerschbaum, Eric Mitchell, John Mitchell, Zulfikar Ramzan, Khawaja Shams, Dawn Song, Ankur Taly, and Diyi Yang. 2023. Identifying and Mitigating the Security Risks of Generative AI. *arXiv preprint arXiv: 2308.14840* (2023).
- [37] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3615–3620. <https://doi.org/10.18653/v1/D19-1371>
- [38] Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, Jeff Clune, Tegan Maharaj, Frank Hutter, Atılım Güneş Baydin, Sheila McIlraith, Qiqi Gao, Ashwin Acharya, David Krueger, Anca Dragan, Philip Torr, Stuart Russell, Daniel Kahneman, Jan Brauner, and Sören Mindermann. 2023. Managing AI Risks in an Era of Rapid Progress. *arXiv preprint arXiv: 2310.17688* (2023).
- [39] Pranjal Bhardwaj, Krishna Yadav, Hind Alsharif, and Rania Anwar Aboalela. 2021. GAN-Based Unsupervised Learning Approach to Generate and Detect Fake News. In *International Conference on Cyber Security, Privacy and Networking*. Springer, 384–396.
- [40] Meghana Moorthy Bhat and Srinivasan Parthasarathy. 2020. How Effectively Can Machines Defend Against Machine-Generated Fake News? An Empirical Study. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*. Association for Computational Linguistics, Online, 48–53. <https://doi.org/10.18653/v1/2020.insights-1.7>
- [41] Amrita Bhattacharjee, Kai Shu, Min Gao, and Huan Liu. 2020. Disinformation in the online information ecosystem: detection, mitigation and challenges. *ArXiv preprint abs/2010.09113* (2020). <https://arxiv.org/abs/2010.09113>
- [42] Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor Detection on Social Media with Bi-Directional Graph Convolutional Networks. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 549–556. <https://aaai.org/ojs/index.php/AAAI/article/view/5393>
- [43] Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. 2023. Safety-Tuned LLaMAs: Lessons From Improving the Safety of Large Language Models that Follow Instructions. *arXiv preprint arXiv: 2309.07875* (2023).
- [44] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy,

- Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kudithipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avaniika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Nieves, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv: Arxiv-2108.07258* (2021).
- [45] Alessandro Bondielli and Francesco Marcelloni. 2019. A survey on fake news and rumour detection techniques. *Information Sciences* 497 (2019), 38–55.
- [46] Sam Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošūtė, Amanda Askell, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, C. McKinnon, C. Olah, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, John Kernion, Jamie Kerr, J. Mueller, Jeff Ladish, J. Landau, Kamal Ndousse, Liane Lovitt, Nelson Elhage, Nicholas Schiefer, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Robin Larson, Sam McCandlish, S. Kundu, Scott Johnston, S. Kravec, Sheer El Showk, Stanislav Fort, Timothy Telleen-Lawton, Tom B. Brown, T. Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Benjamin Mann, and Jared Kaplan. 2022. Measuring Progress on Scalable Oversight for Large Language Models. *ARXIV.ORG* (2022). <https://doi.org/10.48550/arXiv.2211.03540>
- [47] Danilo Brajovic, Niclas Renner, Vincent Philipp Goebels, Philipp Wagner, Benjamin Fresz, Martin Biller, Mara Klueb, Janika Kutz, Jens Neuhüttler, and Marco F. Huber. 2023. Model Reporting for Certifiable AI: A Proposal from Merging EU Regulation into AI Development. *arXiv preprint arXiv: 2307.11525* (2023).
- [48] Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li. 2023. Online Misinformation Video Detection: A Survey. *arXiv preprint arXiv: Arxiv-2302.03242* (2023).
- [49] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv preprint arXiv: Arxiv-2303.12712* (2023).
- [50] Mars Gokturk Buchholz. 2023. Assessing the Effectiveness of GPT-3 in Detecting False Political Statements: A Case Study on the LIAR Dataset. *arXiv preprint arXiv: 2306.08190* (2023).
- [51] Boxi Cao, Hongyu Lin, Xianpei Han, and Le Sun. 2023. The Life Cycle of Knowledge in Big Language Models: A Survey. *arXiv preprint arXiv: 2303.07616* (2023).
- [52] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu, and Lichao Sun. 2023. A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT. *arXiv preprint arXiv: Arxiv-2303.04226* (2023).
- [53] Kevin Matthe Caramancion. 2023. Harnessing the Power of ChatGPT to Decimate Mis/Disinformation: Using ChatGPT for Fake News Detection. In *2023 IEEE World AI IoT Congress (AIIoT)*. IEEE, 0042–0046.
- [54] Fernando Cardoso Durier da Silva, Rafael Vieira, and Ana Cristina Garcia. 2019. Can machines learn to detect fake news? a survey focused on social media. (2019).
- [55] Joseph Carlsmith. 2022. Is Power-Seeking AI an Existential Risk? *arXiv preprint arXiv: 2206.13353* (2022).
- [56] Richard M Carpiano, Timothy Callaghan, Renee DiResta, Noel T Brewer, Chelsea Clinton, Alison P Galvani, Rekha Lakshmanan, Wendy E Parmet, Saad B Omer, Alison M Buttenheim, et al. 2023. Confronting the evolution and expansion of anti-vaccine activism in the USA in the COVID-19 era. *The Lancet* 401, 10380 (2023), 967–970.
- [57] Micah Carroll, Alan Chan, Hal Ashton, and David Krueger. 2023. Characterizing Manipulation from AI Systems. *ARXIV.ORG* (2023). <https://doi.org/10.48550/arXiv.2303.09387>
- [58] Stephen Casper, Jason Lin, Joe Kwon, Gatlen Culp, and Dylan Hadfield-Menell. 2023. Explore, Establish, Exploit: Red Teaming Language Models from Scratch. *arXiv preprint arXiv: 2306.09442* (2023).
- [59] Megha Chakraborty, Khusbu Pahwa, Anku Rani, Adarsh Mahor, Aditya Pakala, Arghya Sarkar, Harshit Dave, Ishan Paul, Janvita Reddy, Preethi Gurumurthy, Ritvik G, Samahriti Mukherjee, Shreyas Chatterjee, Kinjal Sensharma, Dwip Dalal, Suryavardan S, Shreyash Mishra, Parth Patwa, Aman Chadha, Amit Sheth, and Amitava Das. 2023. FACTIFY3M: A Benchmark for Multimodal Fact Verification with Explainability through 5W Question-Answering. *arXiv preprint arXiv: 2306.05523* (2023).
- [60] Souradip Chakraborty, Amrit Singh Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. 2023. On the Possibilities of AI-Generated Text Detection. *arXiv preprint arXiv: Arxiv-2304.04736* (2023).
- [61] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krashennnikov, L. Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, A. Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, and Tegan Maharaj. 2023. Harms from Increasingly Agentic Algorithmic Systems. *ARXIV.ORG* (2023). <https://doi.org/10.48550/arXiv.2302.10329>
- [62] Man-pui Sally Chan, Christopher R Jones, Kathleen Hall Jamieson, and Dolores Albarracín. 2017. Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological science* 28, 11 (2017), 1531–1546.
- [63] Tyler A. Chang and Benjamin K. Bergen. 2023. Language Model Behavior: A Comprehensive Survey. *arXiv preprint arXiv: Arxiv-2303.11504* (2023).
- [64] Chen Chen, Jie Fu, and L. Lyu. 2023. A Pathway Towards Responsible AI Generated Content. *International Joint Conference on Artificial Intelligence* (2023). <https://doi.org/10.48550/arXiv.2303.01325>
- [65] Canyu Chen and Kai Shu. 2023. Can LLM-Generated Misinformation Be Detected? *arXiv preprint arXiv: 2309.13788* (2023).
- [66] Canyu Chen, Haoran Wang, Matthew A. Shapiro, Yunyu Xiao, Fei Wang, and Kai Shu. 2022. Combating Health Misinformation in Social Media: Characterization, Detection, Intervention, and Open Issues. *ARXIV.ORG* (2022). <https://doi.org/10.48550/arXiv.2211.05289>
- [67] Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. Generating Literal and Implied Subquestions to Fact-check Complex Claims. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 3495–3516. <https://aclanthology.org/2022.emnlp-main.229>
- [68] Liang Chen, Yang Deng, Yatao Bian, Zeyu Qin, Bingzhe Wu, Tat-Seng Chua, and Kam-Fai Wong. 2023. Beyond Factuality: A Comprehensive Evaluation of Large Language Models as Knowledge Generators.



- arXiv preprint arXiv: 2310.07289* (2023).
- [69] Lei Chen, Guanying Li, Zhongyu Wei, Yang Yang, Baohua Zhou, Qi Zhang, and Xuanjing Huang. 2022. A Progressive Framework for Role-Aware Rumor Resolution. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2748–2758. <https://aclanthology.org/2022.coling-1.242>
  - [70] Shiqi Chen, Yiran Zhao, Jinghan Zhang, I-Chun Chern, Siyang Gao, Pengfei Liu, and Junxian He. 2023. FELM: Benchmarking Factuality Evaluation of Large Language Models. *arXiv preprint arXiv: 2310.00741* (2023).
  - [71] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2023. AgentVerse: Facilitating Multi-Agent Collaboration and Exploring Emergent Behaviors in Agents. *arXiv preprint arXiv: 2308.10848* (2023).
  - [72] Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Al-abdulmohsin, Piotr Padlewski, Daniel Salz, Xi Xiong, Daniel Vlasic, Filip Pavetic, Keran Rong, Tianli Yu, Daniel Keysers, Xiaohua Zhai, and Radu Soricut. 2023. PaLI-3 Vision Language Models: Smaller, Faster, Stronger. *arXiv preprint arXiv: 2310.09199* (2023).
  - [73] Xiang 'Anthony' Chen, Jeff Burke, Ruofei Du, Matthew K. Hong, Jennifer Jacobs, Philippe Laban, Dingzeyu Li, Nanyun Peng, Karl D. D. Willis, Chien-Sheng Wu, and Bolei Zhou. 2023. Next Steps for Human-Centered Generative AI: A Technical Perspective. *arXiv preprint arXiv: 2306.15774* (2023).
  - [74] Yimin Chen, Niall J Conroy, and Victoria L Rubin. 2015. Misleading online content: recognizing clickbait as "false news". In *Proceedings of the 2015 ACM on workshop on multimodal deception detection*. 15–19.
  - [75] Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-Modal Ambiguity Learning for Multimodal Fake News Detection. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) (WWW '22). Association for Computing Machinery, New York, NY, USA, 2897–2905. <https://doi.org/10.1145/3485447.3511968>
  - [76] Yixuan Chen, Jie Sui, Liang Hu, and Wei Gong. 2019. Attention-Residual Network with CNN for Rumor Detection. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, Wenwu Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu (Eds.). ACM, 1121–1130. <https://doi.org/10.1145/3357384.3357950>
  - [77] Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. 2023. Causal Intervention and Counterfactual Reasoning for Multimodal Fake News Detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 627–638. <https://doi.org/10.18653/v1/2023.acl-long.37>
  - [78] Zhihong Chen, Feng Jiang, Junying Chen, Tiannan Wang, Fei Yu, Guiming Chen, Hongbo Zhang, Juhao Liang, Chen Zhang, Zhiyi Zhang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. Phoenix: Democratizing ChatGPT across Languages. *arXiv preprint arXiv: 2304.10453* (2023).
  - [79] Lu Cheng, Kush R. Varshney, and Huan Liu. 2021. Socially Responsible AI Algorithms: Issues, Purposes, and Challenges. *J. Artif. Intell. Res.* 71 (2021), 1137–1181. <https://doi.org/10.1613/jair.1.12814>
  - [80] Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, Siyin Wang, Xi-angyang Liu, Mozhi Zhang, Junliang He, Mianqiu Huang, Zhangyue Yin, Kai Chen, and Xipeng Qiu. 2023. Evaluating Hallucinations in Chinese Large Language Models. *arXiv preprint arXiv: 2310.03368* (2023).
  - [81] I-Chun Chern, Steffi Chern, Shiqi Chen, Weize Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. FacTool: Factuality Detection in Generative AI - A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios. *arXiv preprint arXiv: 2307.13528* (2023).
  - [82] Tsun-Hin Cheung and Kin-Man Lam. 2023. FactLLaMA: Optimizing Instruction-Following Language Models with External Knowledge for Automated Fact-Checking. *arXiv preprint arXiv: 2309.00240* (2023).
  - [83] Hyewon Choi and Youngjoong Ko. 2021. Using Topic Modeling and Adversarial Neural Networks for Fake News Video Detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (Virtual Event, Queensland, Australia) (CIKM '21). Association for Computing Machinery, New York, NY, USA, 2950–2954. <https://doi.org/10.1145/3459637.3482212>
  - [84] Rajdipa Chowdhury, Sriram Srinivasan, and Lise Getoor. 2020. Joint Estimation of User And Publisher Credibility for Fake News Detection. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux (Eds.). ACM, 1993–1996. <https://doi.org/10.1145/3340531.3412066>
  - [85] Christos Christodoulou, Nikos Salamanos, Pantelitsa Leonidou, Michail Papadakis, and Michael Sirivianos. 2023. Identifying Misinformation on YouTube through Transcript Contextual Analysis with Transformer Models. *arXiv preprint arXiv: 2307.12155* (2023).
  - [86] Samuel Kai Wah Chu, Runbin Xie, and Yanshu Wang. 2021. Cross-language fake news detection. *Data and Information Management* 5, 1 (2021), 100–109.
  - [87] Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2023. A Survey of Chain of Thought Reasoning: Advances, Frontiers and Future. *arXiv preprint arXiv: 2309.15402* (2023).
  - [88] 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* (2023).
  - [89] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational fact checking from knowledge networks. *PloS one* 10, 6 (2015), e0128193.
  - [90] Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. LM vs LM: Detecting Factual Errors via Cross Examination. *arXiv preprint arXiv: 2305.13281* (2023).
  - [91] Andrew Critch and Stuart Russell. 2023. TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI. *arXiv preprint arXiv: 2306.06924* (2023).
  - [92] Jian Cui, Kwanwoo Kim, Seung Ho Na, and Seungwon Shin. 2022. Meta-Path-Based Fake News Detection Leveraging Multi-Level Social Context Information. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (Atlanta, GA, USA) (CIKM '22). Association for Computing Machinery, New York, NY, USA, 325–334. <https://doi.org/10.1145/3511808.3557394>
  - [93] Limeng Cui, Haeseung Seo, Maryam Tabar, Fenglong Ma, Suhang Wang, and Dongwon Lee. 2020. DETERRENT: Knowledge Guided Graph Attention Network for Detecting Healthcare Misinformation. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, Rajesh Gupta, Yan Liu, Jiliang Tang, and B. Aditya Prakash (Eds.). ACM, 492–502. <https://dl.acm.org/doi/10.1145/3394486.3403092>
  - [94] Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-Grained Analysis of Propaganda in News Article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 5636–5646. <https://doi.org/10.18653/v1/D19-1565>

- [95] Arkadipita De, Dibyanayan Bandyopadhyay, Baban Gain, and Asif Ekbal. 2021. A Transformer-Based Approach to Multilingual Fake News Detection in Low-Resource Languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 21, 1, Article 9 (2021), 20 pages. <https://doi.org/10.1145/3472619>
- [96] Luigi De Angelis, Francesco Baglivo, Guglielmo Arzilli, Gaetano Pierpaolo Privitera, Paolo Ferragina, Alberto Eugenio Tozzi, and Caterina Rizzo. 2023. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Frontiers in Public Health* 11 (2023), 1166120.
- [97] Marco Del Tredici and Raquel Fernández. 2020. Words are the Window to the Soul: Language-based User Representations for Fake News Detection. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 5467–5479. <https://doi.org/10.18653/v1/2020.coling-main.477>
- [98] D. Dementieva, Mikhail Kuimov, and A. Panchenko. 2022. Multiverse: Multilingual Evidence for Fake News Detection. *Journal of Imaging* (2022). <https://doi.org/10.3390/jimaging9040077>
- [99] Daryna Dementieva and Alexander Panchenko. 2021. Cross-lingual Evidence Improves Monolingual 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: Student Research Workshop*. Association for Computational Linguistics, Online, 310–320. <https://doi.org/10.18653/v1/2021.acl-srw.32>
- [100] Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. 2023. Jailbreaker: Automated Jailbreak Across Multiple Large Language Model Chatbots. *arXiv preprint arXiv: 2307.08715* (2023).
- [101] Leon Derczynski, Hannah Rose Kirk, Vidhisha Balachandran, Sachin Kumar, Yulia Tsvetkov, M. R. Leiser, and Saif Mohammad. 2023. Assessing Language Model Deployment with Risk Cards. *arXiv preprint arXiv: Arxiv-2303.18190* (2023).
- [102] Erik Derner and Kristina Batistič. 2023. Beyond the Safeguards: Exploring the Security Risks of ChatGPT. *arXiv preprint arXiv: 2305.08005* (2023).
- [103] Matthew R. DeVerna, Harry Yaojun Yan, Kai-Cheng Yang, and Filippo Menczer. 2023. Artificial intelligence is ineffective and potentially harmful for fact checking. *arXiv preprint arXiv: 2308.10800* (2023).
- [104] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- [105] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-Verification Reduces Hallucination in Large Language Models. *arXiv preprint arXiv: 2309.11495* (2023).
- [106] Yingdong Dou, Kai Shu, Congying Xia, Philip S. Yu, and Lichao Sun. 2021. User Preference-Aware Fake News Detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) (SIGIR '21). Association for Computing Machinery, New York, NY, USA, 2051–2055. <https://doi.org/10.1145/3404835.3462990>
- [107] Jiangshu Du, Yingdong Dou, Congying Xia, Limeng Cui, Jing Ma, and Philip S. Yu. 2021. Cross-lingual COVID-19 Fake News Detection. *International Conference on Data Mining Workshops (ICDMW)* (2021). <https://doi.org/10.1109/ICDMW53433.2021.00110>
- [108] Li Du, Yequan Wang, Xingrun Xing, Yiqun Ya, Xiang Li, Xin Jiang, and Xuezhi Fang. 2023. Quantifying and Attributing the Hallucination of Large Language Models via Association Analysis. *arXiv preprint arXiv: 2309.05217* (2023).
- [109] Yibing Du, Antoine Bosselut, and Christopher D. Manning. 2022. Synthetic Disinformation Attacks on Automated Fact Verification Systems. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 10581–10589. <https://ojs.aaai.org/index.php/AAAI/article/view/21302>
- [110] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving Factuality and Reasoning in Language Models through Multiagent Debate. *arXiv preprint arXiv: 2305.14325* (2023).
- [111] Yaqian Dun, Kefei Tu, Chen Chen, Chunyan Hou, and Xiaojie Yuan. 2021. KAN: Knowledge-aware Attention Network for Fake News Detection. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 81–89. <https://ojs.aaai.org/index.php/AAAI/article/view/16080>
- [112] Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. On the Origin of Hallucinations in Conversational Models: Is it the Datasets or the Models?. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 5271–5285. <https://doi.org/10.18653/v1/2022.naacl-main.387>
- [113] El-Mahdi El-Mhamdi, Sadegh Farhadkhani, R. Guerraoui, Nirupam Gupta, L. Hoang, Rafael Pinot, and John Stephan. 2022. On the Impossible Safety of Large AI Models. *ARXIV.ORG* (2022). <https://doi.org/10.48550/arXiv.2209.15259>
- [114] Mohamed Elaraby, Mengyin Lu, Jacob Dunn, Xueying Zhang, Yu Wang, and Shizhu Liu. 2023. Halo: Estimation and Reduction of Hallucinations in Open-Source Weak Large Language Models. *arXiv preprint arXiv: 2308.11764* (2023).
- [115] Ziv Epstein, Antonio Alonso Arechar, and David Rand. 2023. What label should be applied to content produced by generative AI? (2023).
- [116] Martin Fajcik, P. Motlíček, and P. Smrz. 2022. Claim-Dissector: An Interpretable Fact-Checking System with Joint Re-ranking and Veracity Prediction. *Annual Meeting of the Association for Computational Linguistics* (2022). <https://doi.org/10.48550/arXiv.2207.14116>
- [117] Mingyuan Fan, Cen Chen, Chengyu Wang, and Jun Huang. 2023. On the Trustworthiness Landscape of State-of-the-art Generative Models: A Comprehensive Survey. *arXiv preprint arXiv: 2307.16680* (2023).
- [118] Parsa Farinneya, Mohammad Mahdi Abdollah Pour, Sardar Hamidian, and Mona Diab. 2021. Active Learning for Rumor Identification on Social Media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 4556–4565. <https://doi.org/10.18653/v1/2021.findings-emnlp.387>
- [119] Philip Feldman, James R. Foulds, and Shimei Pan. 2023. Trapping LLM Hallucinations Using Tagged Context Prompts. *arXiv preprint arXiv: 2306.06085* (2023).
- [120] Zhangyin Feng, Xiaocheng Feng, Dezhi Zhao, Maojin Yang, and Bing Qin. 2023. Retrieval-Generation Synergy Augmented Large Language Models. *arXiv preprint arXiv: 2310.05149* (2023).
- [121] Patrick Fernandes, Aman Madaan, Emmy Liu, António Farinhas, Pedro Henrique Martins, Amanda Bertsch, José G. C. de Souza, Shuyan Zhou, Tongshuang Sherry Wu, Graham Neubig, and André F. T. Martins. 2023. Bridging the Gap: A Survey on Integrating (Human) Feedback for Natural Language Generation. *ARXIV.ORG* (2023). <https://doi.org/10.48550/arXiv.2305.00955>

- [122] Emilio Ferrara. 2023. GenAI Against Humanity: Nefarious Applications of Generative Artificial Intelligence and Large Language Models. *arXiv preprint arXiv: 2310.00737* (2023).
- [123] Emilio Ferrara. 2023. Social bot detection in the age of ChatGPT: Challenges and opportunities. *First Monday* (2023).
- [124] Anja Folberth, Jutta Jahnke, Jascha Bareis, Carsten Orwat, and Christian Wade. 2022. Tackling problems, harvesting benefits - A systematic review of the regulatory debate around AI. *arXiv preprint arXiv: 2209.05468* (2022).
- [125] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiaowu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv: 2306.13394* (2023).
- [126] Dongqi Fu, Yikun Ban, Hanghang Tong, Ross Maciejewski, and Jingrui He. 2022. DISCO: Comprehensive and Explainable Disinformation Detection. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (Atlanta, GA, USA) (CIKM '22). Association for Computing Machinery, New York, NY, USA, 4848–4852. <https://doi.org/10.1145/3511808.3557202>
- [127] Yi Fung, Kung-Hsiang Huang, Preslav Nakov, and Heng Ji. 2022. The Battlefront of Combating Misinformation and Coping with Media Bias. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Tutorial Abstracts*. Association for Computational Linguistics, Taipei, 28–34. <https://aclanthology.org/2022.aacl-tutorials.5>
- [128] Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avi 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)*. Association for Computational Linguistics, Online, 1683–1698. <https://doi.org/10.18653/v1/2021.acl-long.133>
- [129] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislaw Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. *arXiv preprint arXiv: 2209.07858* (2022).
- [130] Li Gao, Lingyun Song, Jie Liu, Bolin Chen, and Xuequn Shang. 2022. Topology Imbalance and Relation Inauthenticity Aware Hierarchical Graph Attention Networks for Fake News Detection. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 4687–4696. <https://aclanthology.org/2022.coling-1.415>
- [131] Shen Gao, Zhengliang Shi, Minghang Zhu, Bowen Fang, Xin Xin, Pengjie Ren, Zhumin Chen, and Jun Ma. 2023. Confucius: Iterative Tool Learning from Introspection Feedback by Easy-to-Difficult Curriculum. *arXiv preprint arXiv: 2308.14034* (2023).
- [132] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling Large Language Models to Generate Text with Citations. *arXiv preprint arXiv: 2305.14627* (2023).
- [133] Yuan Gao, Xiang Wang, Xiangnan He, Huamin Feng, and Yongdong Zhang. 2022. Rumor Detection with Self-supervised Learning on Texts and Social Graph. *Frontiers Comput. Sci.* (2022). <https://doi.org/10.48550/arXiv.2204.08838>
- [134] Sonal Garg and Dilip Kumar Sharma. 2022. Linguistic features based framework for automatic fake news detection. *Computers & Industrial Engineering* 172 (2022), 108432.
- [135] Valerio La Gatta, Chiyu Wei, Luca Luceri, Francesco Pierri, and Emilio Ferrara. 2023. Retrieving false claims on Twitter during the Russia-Ukraine conflict. *arXiv preprint arXiv: 2303.10121* (2023).
- [136] Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavaš. 2023. mBLIP: Efficient Bootstrapping of Multilingual Vision-LLMs. *arXiv preprint arXiv: 2307.06930* (2023).
- [137] Soumya Suvra Ghosal, Souradip Chakraborty, Jonas Geiping, Furong Huang, Dinesh Manocha, and Amrit Singh Bedi. 2023. Towards Possibilities & Impossibilities of AI-generated Text Detection: A Survey. *arXiv preprint arXiv: 2310.15264* (2023).
- [138] Max Glockner, Yufang Hou, and Iryna Gurevych. 2022. Missing Counter-Evidence Renders NLP Fact-Checking Unrealistic for Misinformation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 5916–5936. <https://aclanthology.org/2022.emnlp-main.397>
- [139] David Glukhov, Ilia Shumailov, Yarin Gal, Nicolas Papernot, and Vardan Papayan. 2023. LLM Censorship: A Machine Learning Challenge or a Computer Security Problem? *arXiv preprint arXiv: 2307.10719* (2023).
- [140] Catalina Goanta, Nikolaos Aletras, Ilias Chalkidis, Sofia Ranchordas, and Gerasimos Spanakis. 2023. Regulation and NLP (RegNLP): Taming Large Language Models. *arXiv preprint arXiv: 2310.05553* (2023).
- [141] Josh A Goldstein, Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. 2023. Can AI Write Persuasive Propaganda? (2023).
- [142] 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.ORG* (2023). <https://doi.org/10.48550/arXiv.2301.04246>
- [143] Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujia Yang, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. ToRA: A Tool-Integrated Reasoning Agent for Mathematical Problem Solving. *arXiv preprint arXiv: 2309.17452* (2023).
- [144] Jon Green, William Hobbs, Stefan McCabe, and David Lazer. 2022. Online engagement with 2020 election misinformation and turnout in the 2021 Georgia runoff election. *Proceedings of the National Academy of Sciences* 119, 34 (2022), e2115900119.
- [145] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, C. Endres, Thorsten Holz, and Mario Fritz. 2023. More than you’ve asked for: A Comprehensive Analysis of Novel Prompt Injection Threats to Application-Integrated Large Language Models. *ARXIV.ORG* (2023). <https://doi.org/10.48550/arXiv.2302.12173>
- [146] Ross Gruetzemacher, Alan Chan, Kevin Frazier, Christy Manning, Štěpán Los, James Fox, José Hernández-Orallo, John Burden, Matija Franklin, Clíodhna Ní Ghuidhir, Mark Bailey, Daniel Eth, Toby Pilditch, and Kyle Kilian. 2023. An International Consortium for Evaluations of Societal-Scale Risks from Advanced AI. *arXiv preprint arXiv: 2310.14455* (2023).
- [147] Nuno M. Guerreiro, Duarte Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in Large Multilingual Translation Models. *arXiv preprint arXiv: 2303.16104* (2023).
- [148] Andrew M Guess and Benjamin A Lyons. 2020. Misinformation, disinformation, and online propaganda. *Social media and democracy: The state of the field, prospects for reform* 10 (2020).
- [149] Bin Guo, Yasan Ding, Yueheng Sun, Shuai Ma, and Ke Li. 2019. The Mass, Fake News, and Cognition Security. *Frontiers of Computer Science* (2019). <https://doi.org/10.1007/s11704-020-9256-0>

- [150] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. *arXiv preprint arXiv: Arxiv-2301.07597* (2023).
- [151] Danhuai Guo, Huixuan Chen, Ruoling Wu, and Yangang Wang. 2023. AIGC Challenges and Opportunities Related to Public Safety: A Case Study of ChatGPT. *Journal of Safety Science and Resilience* (2023).
- [152] Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. 2023. Evaluating Large Language Models: A Comprehensive Survey. *arXiv preprint arXiv: 2310.19736* (2023).
- [153] Ashim Gupta and Vivek Srikumar. 2021. X-Fact: A New Benchmark Dataset for Multilingual Fact Checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, Online, 675–682. <https://doi.org/10.18653/v1/2021.acl-short.86>
- [154] Prakhhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. DialFact: A Benchmark for Fact-Checking in Dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 3785–3801. <https://doi.org/10.18653/v1/2022.acl-long.263>
- [155] Vipin Gupta, Rina Kumari, Nischal Ashok, Tirthankar Ghosal, and Asif Ekbal. 2022. MMM: An Emotion and Novelty-aware Approach for Multilingual Multimodal Misinformation Detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2022*. Association for Computational Linguistics, Online only, 464–477. <https://aclanthology.org/2022.findings-acl.43>
- [156] P. Hacker. 2023. Sustainable AI Regulation. *Social Science Research Network* (2023). <https://doi.org/10.48550/arXiv.2306.00292>
- [157] Philipp Hacker, Andreas Engel, and Marco Mauer. 2023. Regulating ChatGPT and other Large Generative AI Models. *arXiv preprint arXiv: Arxiv-2302.02337* (2023).
- [158] Gillian K. Hadfield and Jack Clark. 2023. Regulatory Markets: The Future of AI Governance. *arXiv preprint arXiv: 2304.04914* (2023).
- [159] Samar Haider, Luca Luceri, A. Deb, Adam Badawy, Nanyun Peng, and Emilio Ferrara. 2020. Detecting Social Media Manipulation in Low-Resource Languages. *The Web Conference* (2020). <https://doi.org/10.1145/3543873.3587615>
- [160] Ahmed Abdeen Hamed. 2023. Improving Detection of ChatGPT-Generated Fake Science Using Real Publication Text: Introducing xFakeBibs a Supervised Learning Network Algorithm. (2023).
- [161] Hicham Hammouchi and Mounir Ghogho. 2022. Evidence-Aware Multilingual Fake News Detection. *IEEE Access* 10 (2022), 116808–116818. <https://doi.org/10.1109/ACCESS.2022.3220690>
- [162] Hans W. A. Hanley and Zakir Durumeric. 2023. Machine-Made Media: Monitoring the Mobilization of Machine-Generated Articles on Misinformation and Mainstream News Websites. *arXiv preprint arXiv: 2305.09820* (2023).
- [163] Fatima Haouari. 2022. Evidence-Based Early Rumor Verification in Social Media. In *European Conference on Information Retrieval*. Springer, 496–504.
- [164] Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. A Survey on Stance Detection for Mis- and Disinformation Identification. In *Findings of the Association for Computational Linguistics: NAACL 2022*. Association for Computational Linguistics, Seattle, United States, 1259–1277. <https://doi.org/10.18653/v1/2022.findings-naacl.94>
- [165] Katrin Hartwig, Frederic Doell, and Christian Reuter. 2023. The Landscape of User-centered Misinformation Interventions - A Systematic Literature Review. *arXiv preprint arXiv: Arxiv-2301.06517* (2023).
- [166] Bing He, Mustaque Ahamad, and Srijan Kumar. 2023. Reinforcement Learning-Based Counter-Misinformation Response Generation: A Case Study of COVID-19 Vaccine Misinformation. In *Proceedings of the ACM Web Conference 2023* (Austin, TX, USA) (WWW '23). Association for Computing Machinery, New York, NY, USA, 2698–2709. <https://doi.org/10.1145/3543507.3583388>
- [167] Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2023. A Survey of Large Language Models for Healthcare: from Data, Technology, and Applications to Accountability and Ethics. *arXiv preprint arXiv: 2310.05694* (2023).
- [168] Qianyu He, Jie Zeng, Wenhao Huang, Lina Chen, Jin Xiao, Qianxi He, Xunzhe Zhou, Lida Chen, Xintao Wang, Yuncheng Huang, Haoning Ye, Zihan Li, Shisong Chen, Yikai Zhang, Zhouhong Gu, Jiaqing Liang, and Yanghua Xiao. 2023. Can Large Language Models Understand Real-World Complex Instructions? *arXiv preprint arXiv: 2309.09150* (2023).
- [169] Zhenyu He, Ce Li, Fan Zhou, and Yi Yang. 2021. Rumor Detection on Social Media with Event Augmentations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) (SIGIR '21). Association for Computing Machinery, New York, NY, USA, 2020–2024. <https://doi.org/10.1145/3404835.3463001>
- [170] 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* (2023).
- [171] Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A. Lemley, and Percy Liang. 2023. Foundation Models and Fair Use. *arXiv preprint arXiv: Arxiv-2303.15715* (2023).
- [172] Peter Henderson, E. Mitchell, Christopher D. Manning, Dan Jurafsky, and Chelsea Finn. 2022. Self-Destructing Models: Increasing the Costs of Harmful Dual Uses of Foundation Models. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (2022). <https://doi.org/10.1145/3600211.3604690>
- [173] Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. 2023. An Overview of Catastrophic AI Risks. *arXiv preprint arXiv: 2306.12001* (2023).
- [174] Da Silva Gameiro Henrique, Andrei Kucharavy, and Rachid Guerraoui. 2023. Stochastic Parrots Looking for Stochastic Parrots: LLMs are Easy to Fine-Tune and Hard to Detect with other LLMs. *arXiv preprint arXiv: Arxiv-2304.08968* (2023).
- [175] Steffen Herbold, Annette Hautli-Janisz, Ute Heuer, Zlata Kikteva, and Alexander Trautsch. 2023. AI, write an essay for me: A large-scale comparison of human-written versus ChatGPT-generated essays. *arXiv preprint arXiv: Arxiv-2304.14276* (2023).
- [176] Lewis Ho, Joslyn Barnhart, Robert Trager, Yoshua Bengio, Miles Brundage, Allison Carnegie, Rumman Chowdhury, Allan Dafoe, Gillian Hadfield, Margaret Levi, and Duncan Sndal. 2023. International Institutions for Advanced AI. *arXiv preprint arXiv: 2307.04699* (2023).
- [177] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [178] Emma Hoes, Sacha Altay, and Juan Bermeo. 2023. Leveraging ChatGPT for Efficient Fact-Checking. (2023).
- [179] Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, and Chenglin Wu. 2023. MetaGPT: Meta Programming for Multi-Agent Collaborative Framework. *arXiv preprint arXiv: 2308.00352* (2023).
- [180] Benjamin D. Horne and Sibel Adali. 2017. This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News. *arXiv:1703.09398* [cs.SI]
- [181] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 159–166.
- [182] Abe Bohan Hou, Jingyu Zhang, Tianxing He, Yichen Wang, Yung-Sung Chuang, Hongwei Wang, Lingfeng Shen, Benjamin Van Durme,

- Daniel Khashabi, and Yulia Tsvetkov. 2023. SemStamp: A Semantic Watermark with Paraphrastic Robustness for Text Generation. *arXiv preprint arXiv: 2310.03991* (2023).
- [183] Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2023. Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection. *arXiv preprint arXiv: 2309.12247* (2023).
- [184] Beizhe Hu, Qiang Sheng, Juan Cao, Yongchun Zhu, Danding Wang, Zhengjia Wang, and Zhiwei Jin. 2023. Learn over Past, Evolve for Future: Forecasting Temporal Trends for Fake News Detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*. Association for Computational Linguistics, Toronto, Canada, 116–125. <https://doi.org/10.18653/v1/2023.acl-industry.13>
- [185] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net. <https://openreview.net/forum?id=nZeVKeeFYf9>
- [186] Linmei Hu, Zeyi Liu, Ziwang Zhao, Lei Hou, Liqiang Nie, and Juanzi Li. 2023. A Survey of Knowledge Enhanced Pre-Trained Language Models. *IEEE Transactions on Knowledge and Data Engineering* (2023).
- [187] Linmei Hu, Tianchi Yang, Luhao Zhang, Wanjun Zhong, Duyu Tang, Chuan Shi, Nan Duan, and Ming Zhou. 2021. Compare to The Knowledge: Graph Neural Fake News Detection with External Knowledge. 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)*. Association for Computational Linguistics, Online, 754–763. <https://doi.org/10.18653/v1/2021.acl-long.62>
- [188] Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S. Yu, and Zhijiang Guo. 2023. Do Large Language Models Know about Facts? *arXiv preprint arXiv: 2310.05177* (2023).
- [189] Xuming Hu, Zhijiang Guo, Junzhe Chen, Lijie Wen, and Philip S. Yu. 2023. MR2: A Benchmark for Multimodal Retrieval-Augmented Rumor Detection in Social Media. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Taipei, Taiwan) (SIGIR '23). Association for Computing Machinery, New York, NY, USA, 2901–2912. <https://doi.org/10.1145/3539618.3591896>
- [190] Xuming Hu, Zhijiang Guo, Guanyu Wu, Lijie Wen, and Philip S. Yu. 2023. Give Me More Details: Improving Fact-Checking with Latent Retrieval. *arXiv preprint arXiv: 2305.16128* (2023).
- [191] Jie Huang and K. Chang. 2022. Towards Reasoning in Large Language Models: A Survey. *Annual Meeting of the Association for Computational Linguistics* (2022). <https://doi.org/10.48550/arXiv.2212.10403>
- [192] Jie Huang and Kevin Chen-Chuan Chang. 2023. Citation: A Key to Building Responsible and Accountable Large Language Models. *arXiv preprint arXiv: 2307.02185* (2023).
- [193] Kung-Hsiang Huang, Kathleen McKeown, Preslav Nakov, Yejin Choi, and Heng Ji. 2022. Faking Fake News for Real Fake News Detection: Propaganda-loaded Training Data Generation. *arXiv preprint arXiv: 2203.05386* (2022).
- [194] Kung-Hsiang Huang, ChengXiang Zhai, and Heng Ji. 2022. CON-CRETE: Improving Cross-lingual Fact-checking with Cross-lingual Retrieval. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 1024–1035. <https://aclanthology.org/2022.coling-1.86>
- [195] Linan Huang and Quanyan Zhu. 2023. An Introduction of System-Scientific Approaches to Cognitive Security. *arXiv preprint arXiv: 2301.05920* (2023).
- [196] Xiaowei Huang, Wenjie Ruan, Wei Huang, Gaojie Jin, Yi Dong, Changshun Wu, Saddek Bensalem, Ronghui Mu, Yi Qi, Xingyu Zhao, Kaiwen Cai, Yanghao Zhang, Sihao Wu, Peipei Xu, Dengyu Wu, Andre Freitas, and Mustafa A. Mustafa. 2023. A Survey of Safety and Trustworthiness of Large Language Models through the Lens of Verification and Validation. *arXiv preprint arXiv: 2305.11391* (2023).
- [197] Yinqiu Huang, Min Gao, Jia Wang, and Kai Shu. 2021. Dafd: Domain adaptation framework for fake news detection. In *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part I 28*. Springer, 305–316.
- [198] Yinqiu Huang, Min Gao, Jia Wang, Junwei Yin, Kai Shu, Qilin Fan, and Junhao Wen. 2023. Meta-prompt based learning for low-resource false information detection. *Information Processing & Management* 60, 3 (2023), 103279.
- [199] Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. 2023. Catastrophic Jailbreak of Open-source LLMs via Exploiting Generation. *arXiv:2310.06987* [cs.CL]
- [200] Yue Huang, Jiawen Shi, Yuan Li, Chenrui Fan, Siyuan Wu, Qihui Zhang, Yixin Liu, Pan Zhou, Yao Wan, Neil Zhenqiang Gong, and Lichao Sun. 2023. MetaTool Benchmark for Large Language Models: Deciding Whether to Use Tools and Which to Use. *arXiv preprint arXiv: 2310.03128* (2023).
- [201] Yuheng Huang, Jiayang Song, Zhijie Wang, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2023. Look Before You Leap: An Exploratory Study of Uncertainty Measurement for Large Language Models. *arXiv preprint arXiv: 2307.10236* (2023).
- [202] Yue Huang and Lichao Sun. 2023. Harnessing the Power of ChatGPT in Fake News: An In-Depth Exploration in Generation, Detection and Explanation. *arXiv preprint arXiv: 2310.05046* (2023).
- [203] Yue Huang, Qihui Zhang, Philip S. Y, and Lichao Sun. 2023. Trust-GPT: A Benchmark for Trustworthy and Responsible Large Language Models. *arXiv preprint arXiv: 2306.11507* (2023).
- [204] Zhen Huang, Zhilong Lv, Xiaoyun Han, Binyang Li, Menglong Lu, and Dongsheng Li. 2022. Social Bot-Aware Graph Neural Network for Early Rumor Detection. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 6680–6690. <https://aclanthology.org/2022.coling-1.580>
- [205] Umar Iqbal, Tadayoshi Kohno, and Franziska Roesner. 2023. LLM Platform Security: Applying a Systematic Evaluation Framework to OpenAI’s ChatGPT Plugins. *arXiv preprint arXiv: 2309.10254* (2023).
- [206] 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 (2020), 1–20.
- [207] J Jaiswal, C LoSchiavo, and DC Perlman. 2020. Disinformation, misinformation and inequality-driven mistrust in the time of COVID-19: lessons unlearned from AIDS denialism. *AIDS and Behavior* 24, 10 (2020), 2776–2780.
- [208] Ujun Jeong, Kaize Ding, Lu Cheng, Ruocheng Guo, Kai Shu, and Huan Liu. 2022. Nothing stands alone: Relational fake news detection with hypergraph neural networks. In *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, 596–605.
- [209] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O’Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. 2023. AI Alignment: A Comprehensive Survey. *arXiv preprint arXiv: 2310.19852* (2023).
- [210] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, D. Su, Yan Xu, Etsuko Ishii, Yejin Bang, Wenliang Dai, Andrea Madotto, and Pascale Fung. 2022. Survey of Hallucination in Natural Language Generation. *Acm*



- Computing Surveys* (2022). <https://doi.org/10.1145/3571730>
- [211] Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards Mitigating Hallucination in Large Language Models via Self-Reflection. *arXiv preprint arXiv: 2310.06271* (2023).
  - [212] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint arXiv: 2310.06825* (2023).
  - [213] Bohan Jiang, Zhen Tan, Ayushi Nirmal, and Huan Liu. 2023. Disinformation Detection: An Evolving Challenge in the Age of LLMs. *arXiv preprint arXiv: 2309.15847* (2023).
  - [214] Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active Retrieval Augmented Generation. *arXiv preprint arXiv: 2305.06983* (2023).
  - [215] Yiqiao Jin, Xiting Wang, Ruichao Yang, Yizhou Sun, Wei Wang, Hao Liao, and Xing Xie. 2022. Towards Fine-Grained Reasoning for Fake News Detection. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 5746–5754. <https://ojs.aaai.org/index.php/AAAI/article/view/20517>
  - [216] Pica Johansson, Florence Enock, Scott Hale, Bertie Vidgen, Cassidy Bereskin, Helen Margetts, and Jonathan Bright. 2022. How can we combat online misinformation? A systematic overview of current interventions and their efficacy. *arXiv preprint arXiv: 2212.11864* (2022).
  - [217] Prerna Juneja, M. Bhuiyan, and Tanushree Mitra. 2023. Assessing enactment of content regulation policies: A post hoc crowd-sourced audit of election misinformation on YouTube. *International Conference on Human Factors in Computing Systems* (2023). <https://doi.org/10.1145/3544548.3580846>
  - [218] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A Convolutional Neural Network for Modelling Sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Baltimore, Maryland, 655–665. <https://doi.org/10.3115/v1/P14-1062>
  - [219] Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. FakeBERT: Fake News Detection in Social Media with a BERT-Based Deep Learning Approach. *Multimedia Tools Appl.* 80, 8 (2021), 11765–11788. <https://doi.org/10.1007/s11042-020-10183-2>
  - [220] Daniel Kang, Xuechen Li, I. Stoica, Carlos Guestrin, M. Zaharia, and Tatsunori Hashimoto. 2023. Exploiting Programmatic Behavior of LLMs: Dual-Use Through Standard Security Attacks. *ARXIV.ORG* (2023). <https://doi.org/10.48550/arXiv.2302.05733>
  - [221] Mert Karabacak and Konstantinos Margetis. 2023. Embracing Large Language Models for Medical Applications: Opportunities and Challenges. *Cureus* 15, 5 (2023).
  - [222] Hamid Karimi and Jiliang Tang. 2019. Learning Hierarchical Discourse-level Structure for Fake News Detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 3432–3442. <https://doi.org/10.18653/v1/N19-1347>
  - [223] Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T. Hancock. 2023. Working With AI to Persuade: Examining a Large Language Model’s Ability to Generate Pro-Vaccination Messages. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 116 (2023), 29 pages. <https://doi.org/10.1145/3579592>
  - [224] Krishnaram Kenthapadi, Himabindu Lakkaraju, and Nazneen Rajani. 2023. Generative AI Meets Responsible AI: Practical Challenges and Opportunities. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (Long Beach, CA, USA) (KDD ’23)*. Association for Computing Machinery, New York, NY, USA, 5805–5806. <https://doi.org/10.1145/3580305.3599557>
  - [225] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: Multimodal Variational Autoencoder for Fake News Detection. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia (Eds.). ACM, 2915–2921. <https://doi.org/10.1145/3308558.3313552>
  - [226] Ling Min Serena Khoo, Hai Leong Chieu, Zhong Qian, and Jing Jiang. 2020. Interpretable Rumor Detection in Microblogs by Attending to User Interactions. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 8783–8790. <https://aaai.org/ojs/index.php/AAAI/article/view/6405>
  - [227] Jagdish Khubchandani and Yilda Macias. 2021. COVID-19 vaccination hesitancy in Hispanics and African-Americans: A review and recommendations for practice. *Brain, behavior, & immunity-health* 15 (2021), 100277.
  - [228] Celeste Kidd and Abeba Birhane. 2023. How AI can distort human beliefs. *Science* 380, 6651 (2023), 1222–1223.
  - [229] Jiho Kim, Sungjin Park, Yeonsu Kwon, Yohan Jo, James Thorne, and E. Choi. 2023. FactKG: Fact Verification via Reasoning on Knowledge Graphs. *Annual Meeting of the Association for Computational Linguistics* (2023). <https://doi.org/10.48550/arXiv.2305.06590>
  - [230] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A Watermark for Large Language Models. *arXiv preprint arXiv: Arxiv-2301.10226* (2023).
  - [231] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). <https://openreview.net/forum?id=e2TBb5y0yFf>
  - [232] Sai Koneru, Jian Wu, and Sarah Rajtmajer. 2023. Can Large Language Models Discern Evidence for Scientific Hypotheses? Case Studies in the Social Sciences. *arXiv preprint arXiv: 2309.06578* (2023).
  - [233] Ziyi Kou, Lanyu Shang, Yang Zhang, Zhenrui Yue, Huimin Zeng, and Dong Wang. 2022. Crowd, Expert & AI: A Human-AI Interactive Approach Towards Natural Language Explanation Based COVID-19 Misinformation Detection. In *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*. 5087–5093.
  - [234] Canasai Kruengkrai, Junichi Yamagishi, and Xin Wang. 2021. A Multi-Level Attention Model for Evidence-Based Fact Checking. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, Online, 2447–2460. <https://doi.org/10.18653/v1/2021.findings-acl.217>
  - [235] Rohith Kudithipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. 2023. Robust Distortion-free Watermarks for Language Models. *arXiv preprint arXiv: 2307.15593* (2023).
  - [236] Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Soheil Feizi, and Hima Lakkaraju. 2023. Certifying LLM Safety against Adversarial Prompting. *arXiv preprint arXiv: 2309.02705* (2023).
  - [237] Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastopoulos, and Yulia Tsvetkov. 2023. Language Generation Models Can Cause Harm: So What Can We Do About It? An Actionable Survey. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Dubrovnik, Croatia, 3299–3321.

<https://aclanthology.org/2023.eacl-main.241>

- [238] Srijan Kumar and Neil Shah. 2018. False information on web and social media: A survey. *ArXiv preprint abs/1804.08559* (2018). <https://arxiv.org/abs/1804.08559>
- [239] Srijan Kumar, Robert West, and Jure Leskovec. 2016. Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, Jacqueline Bourdeau, Jim Hendler, Roger Nkambou, Ian Horrocks, and Ben Y. Zhao (Eds.). ACM, 591–602. <https://doi.org/10.1145/2872427.2883085>
- [240] Tharindu Kumarage, Amrita Bhattacharjee, Djordje Padejski, Kristy Roschke, Dan Gillmor, Scott Ruston, Huan Liu, and Joshua Garland. 2023. J-Guard: Journalism Guided Adversarially Robust Detection of AI-generated News. *arXiv preprint arXiv: 2309.03164* (2023).
- [241] Tharindu Kumarage, Joshua Garland, Amrita Bhattacharjee, Kirill Trapeznikov, Scott Ruston, and Huan Liu. 2023. Stylometric Detection of AI-Generated Text in Twitter Timelines. *arXiv preprint arXiv: Arxiv-2303.03697* (2023).
- [242] Sandipan Kundu, Yuntao Bai, Saurav Kadavath, Amanda Askell, Andrew Callahan, Anna Chen, Anna Goldie, Avital Balwit, Azalia Mirhoseini, Brayden McLean, Catherine Olsson, Cassie Evraets, Eli Tran-Johnson, Esin Durmus, Ethan Perez, Jackson Kernion, Jamie Kerr, Kamal Ndousse, Karina Nguyen, Nelson Elhage, Newton Cheng, Nicholas Schiefer, Nova DasSarma, Oliver Rausch, Robin Larson, Shannon Yang, Shauna Kravec, Timothy Telleen-Lawton, Thomas I. Liao, Tom Henighan, Tristan Hume, Zac Hatfield-Dodds, Sören Mindermann, Nicholas Joseph, Sam McCandlish, and Jared Kaplan. 2023. Specific versus General Principles for Constitutional AI. *arXiv preprint arXiv: 2310.13798* (2023).
- [243] Laida Kushnareva, Daniil Cherniavskii, Vladislav Mikhailov, Ekaterina Artemova, Serguei Barannikov, Alexander Bernstein, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. 2021. Artificial Text Detection via Examining the Topology of Attention Maps. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 635–649. <https://doi.org/10.18653/v1/2021.emnlp-main.50>
- [244] Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. *arXiv preprint arXiv: Arxiv-2304.05613* (2023).
- [245] Hussain S Lalani, Renée DiResta, Richard J Baron, and David Scales. 2023. Addressing Viral Medical Rumors and False or Misleading Information. *Annals of Internal Medicine* 176, 8 (2023), 1113–1120.
- [246] An Lao, Chongyang Shi, and Yayi Yang. 2021. Rumor Detection with Field of Linear and Non-Linear Propagation. In *Proceedings of the Web Conference 2021 (Ljubljana, Slovenia) (WWW '21)*. Association for Computing Machinery, New York, NY, USA, 3178–3187. <https://doi.org/10.1145/3442381.3450016>
- [247] Raz Lapid, Ron Langberg, and Moshe Sipper. 2023. Open Sesame! Universal Black Box Jailbreaking of Large Language Models. *arXiv preprint arXiv: 2309.01446* (2023).
- [248] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.
- [249] Thai Le, Suhang Wang, and Dongwon Lee. 2020. MALCOM: Generating Malicious Comments to Attack Neural Fake News Detection Models. In *20th IEEE International Conference on Data Mining, ICDM 2020, Sorrento, Italy, November 17–20, 2020*, Claudia Plant, Haixun Wang, Alfredo Cuzzocrea, Carlo Zaniolo, and Xindong Wu (Eds.). IEEE, 282–291. <https://doi.org/10.1109/ICDM50108.2020.00037>
- [250] Deokjae Lee, JunYeong Lee, Jung-Woo Ha, Jin-Hwa Kim, Sang-Woo Lee, Hwaran Lee, and Hyun Oh Song. 2023. Query-Efficient Black-Box Red Teaming via Bayesian Optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 11551–11574. <https://doi.org/10.18653/v1/2023.acl-long.646>
- [251] Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. 2021. Towards Few-shot Fact-Checking via Perplexity. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 1971–1981. <https://doi.org/10.18653/v1/2021.naacl-main.158>
- [252] Nayeon Lee, Belinda Z. Li, Sinong Wang, Pascale Fung, Hao Ma, Wen-tau Yih, and Madian Khabsa. 2021. On Unifying Misinformation Detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 5479–5485. <https://doi.org/10.18653/v1/2021.naacl-main.432>
- [253] Nayeon Lee, Wei Ping, Peng Xu, M. Patwary, M. Shoenybi, and Bryan Catanzaro. 2022. Factuality Enhanced Language Models for Open-Ended Text Generation. *ARXIV.ORG* (2022). <https://doi.org/10.48550/arXiv.2206.04624>
- [254] Taehyun Lee, Seokhee Hong, Jaewoo Ahn, Ilgee Hong, Hwaran Lee, Sangdoo Yun, Jamin Shin, and Gunhee Kim. 2023. Who Wrote this Code? Watermarking for Code Generation. *arXiv preprint arXiv: 2305.15060* (2023).
- [255] Deren Lei, Yaxi Li, Mengya Hu, Mingyu Wang, Vincent Yun, Emily Ching, and Eslam Kamal. 2023. Chain of Natural Language Inference for Reducing Large Language Model Ungrounded Hallucinations. *arXiv preprint arXiv: 2310.03951* (2023).
- [256] 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* (2023).
- [257] Or Levi, Pedram Hosseini, Mona Diab, and David Broniatowski. 2019. Identifying Nuances in Fake News vs. Satire: Using Semantic and Linguistic Cues. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*. Association for Computational Linguistics, Hong Kong, China, 31–35. <https://doi.org/10.18653/v1/D19-5004>
- [258] Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest* 13, 3 (2012), 106–131.
- [259] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension. *arXiv preprint arXiv: 2307.16125* (2023).
- [260] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. 2023. Multimodal Foundation Models: From Specialists to General-Purpose Assistants. *arXiv preprint arXiv: 2309.10020* (2023).
- [261] Chen Li, Hao Peng, Jianxin Li, Lichao Sun, Lingjuan Lyu, Lihong Wang, Philip S. Yu, and Lifang He. 2022. Joint Stance and Rumor Detection in Hierarchical Heterogeneous Graph. *IEEE Transactions on Neural Networks and Learning Systems* 33, 6 (2022), 2530–2542. <https://doi.org/10.1109/TNNLS.2021.3114027>
- [262] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative Agents for "Mind" Exploration of Large Scale Language Model Society. *arXiv preprint arXiv: 2303.17760* (2023).
- [263] Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, and Yangqiu Song. 2023. Multi-step Jailbreaking Privacy Attacks on ChatGPT. *ARXIV.ORG* (2023). <https://doi.org/10.48550/arXiv.2304.05197>

- [264] Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. 2022. A Survey on Retrieval-Augmented Text Generation. *arXiv preprint arXiv: 2202.01110* (2022).
- [265] Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. *arXiv preprint arXiv: 2305.11747* (2023).
- [266] Jianning Li, Amin Dada, Jens Kleesiek, and Jan Egger. 2023. ChatGPT in Healthcare: A Taxonomy and Systematic Review. *medRxiv* (2023), 2023–03.
- [267] Jiawen Li, Shiwen Ni, and Hung-Yu Kao. 2021. Meet The Truth: Leverage Objective Facts and Subjective Views for Interpretable Rumor Detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, Online, 705–715. <https://doi.org/10.18653/v1/2021.findings-acl.63>
- [268] Jiawen Li, Yudianto Sujana, and Hung-Yu Kao. 2020. Exploiting Microblog Conversation Structures to Detect Rumors. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 5420–5429. <https://doi.org/10.18653/v1/2020.coling-main.473>
- [269] Ke Li, Bin Guo, Siyuan Ren, and Zhiwen Yu. 2022. AdaDeBunk: An Efficient and Reliable Deep State Space Model for Adaptive Fake News Early Detection. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (Atlanta, GA, USA) (CIKM '22)*. Association for Computing Machinery, New York, NY, USA, 1156–1165. <https://doi.org/10.1145/3511808.3557227>
- [270] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. *arXiv preprint arXiv: 2306.03341* (2023).
- [271] Miaoran Li, Baolin Peng, and Zhu Zhang. 2023. Self-Checker: Plug-and-Play Modules for Fact-Checking with Large Language Models. *arXiv preprint arXiv: 2305.14623* (2023).
- [272] Quanzhi Li, Qiong Zhang, and Luo Si. 2019. Rumor Detection by Exploiting User Credibility Information, Attention and Multi-task Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1173–1179. <https://doi.org/10.18653/v1/P19-1113>
- [273] Quanzhi Li, Qiong Zhang, Luo Si, and Yingchi Liu. 2019. Rumor Detection on Social Media: Datasets, Methods and Opportunities. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*. Association for Computational Linguistics, Hong Kong, China, 66–75. <https://doi.org/10.18653/v1/D19-5008>
- [274] Qifei Li and Wangchunshu Zhou. 2020. Connecting the Dots Between Fact Verification and Fake News Detection. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 1820–1825. <https://doi.org/10.18653/v1/2020.coling-main.165>
- [275] Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, and Jie Yu. 2023. PMET: Precise Model Editing in a Transformer. *arXiv preprint arXiv: 2308.08742* (2023).
- [276] Xinyi Li, Yongfeng Zhang, and Edward C. Malthouse. 2023. A Preliminary Study of ChatGPT on News Recommendation: Personalization, Provider Fairness, Fake News. *arXiv preprint arXiv: 2306.10702* (2023).
- [277] Yichuan Li, Kyumin Lee, Nima Kordzadeh, Brenton Faber, Cameron Fiddes, Elaine Chen, and Kai Shu. 2021. Multi-source domain adaptation with weak supervision for early fake news detection. In *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 668–676.
- [278] Yafu Li, Quintong Li, Leyang Cui, Wei Bi, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2023. Deepfake Text Detection in the Wild. *arXiv preprint arXiv: 2305.13242* (2023).
- [279] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research* (2023). <https://openreview.net/forum?id=iO4LZibEqW> Featured Certification, Expert Certification.
- [280] Hao Liao, Jiahao Peng, Zhanyi Huang, Wei Zhang, Guanghua Li, Kai Shu, and Xing Xie. 2023. MUSER: A Multi-Step Evidence Retrieval Enhancement Framework for Fake News Detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4461–4472.
- [281] Q. Vera Liao and Jennifer Wortman Vaughan. 2023. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *arXiv preprint arXiv: 2306.01941* (2023).
- [282] Wenxiong Liao, Zhengliang Liu, Haixing Dai, Shaochen Xu, Zihao Wu, Yiyang Zhang, Xiaoke Huang, Dajiang Zhu, Hongmin Cai, Tianming Liu, and Xiang Li. 2023. Differentiate ChatGPT-generated and Human-written Medical Texts. *arXiv preprint arXiv: Arxiv-2304.11567* (2023).
- [283] Anders Edelbo Lillie and Emil Refsgaard Middelboe. 2019. Fake news detection using stance classification: A survey. *ArXiv preprint abs/1907.00181* (2019). <https://arxiv.org/abs/1907.00181>
- [284] Hongzhan Lin, Jing Ma, Liangliang Chen, Zhiwei Yang, Mingfei Cheng, and Chen Guang. 2022. Detect Rumors in Microblog Posts for Low-Resource Domains via Adversarial Contrastive Learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*. Association for Computational Linguistics, Seattle, United States, 2543–2556. <https://doi.org/10.18653/v1/2022.findings-naacl.194>
- [285] Hongzhan Lin, Jing Ma, Mingfei Cheng, Zhiwei Yang, Liangliang Chen, and Guang Chen. 2021. Rumor Detection on Twitter with Claim-Guided Hierarchical Graph Attention Networks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 10035–10047. <https://doi.org/10.18653/v1/2021.emnlp-main.786>
- [286] Hongzhan Lin, Pengyao Yi, Jing Ma, Haiyun Jiang, Ziyang Luo, Shuming Shi, and Ruifang Liu. 2023. Zero-shot rumor detection with propagation structure via prompt learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 5213–5221.
- [287] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 3214–3252. <https://doi.org/10.18653/v1/2022.acl-long.229>
- [288] Chengyuan Liu, Fubang Zhao, Lizhi Qing, Yangyang Kang, Changlong Sun, Kun Kuang, and Fei Wu. 2023. A Chinese Prompt Attack Dataset for LLMs with Evil Content. *arXiv preprint arXiv: 2309.11830* (2023).
- [289] Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023. HallusionBench: You See What You Think? Or You Think What You See? An Image-Context Reasoning Benchmark Challenging for GPT-4V(ision), LLaVA-1.5, and Other Multi-modality Models. *arXiv:2310.14566 [cs.CV]*

- [290] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023. Aligning Large Multi-Modal Model with Robust Instruction Tuning. *ArXiv preprint abs/2306.14565* (2023). <https://arxiv.org/abs/2306.14565>
- [291] Fuxiao Liu, Yaser Yacoob, and Abhinav Shrivastava. 2023. COVID-VTS: Fact Extraction and Verification on Short Video Platforms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Dubrovnik, Croatia, 178–188. <https://aclanthology.org/2023.eacl-main.14>
- [292] Haoyang Liu, Maheep Chaudhary, and Haohan Wang. 2023. Towards Trustworthy and Aligned Machine Learning: A Data-centric Survey with Causality Perspectives. *arXiv preprint arXiv: 2307.16851* (2023).
- [293] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. *ARXIV.ORG* (2023). <https://doi.org/10.48550/arXiv.2304.08485>
- [294] Hui Liu, Wenya Wang, and Hao Li. 2023. Interpretable Multimodal Misinformation Detection with Logic Reasoning. *Annual Meeting of the Association for Computational Linguistics* (2023). <https://doi.org/10.48550/arXiv.2305.05964>
- [295] Haochen Liu, Yiqi Wang, Wenqi Fan, Xiaorui Liu, Yaxin Li, Shaili Jain, Anil K. Jain, and Jiliang Tang. 2021. Trustworthy AI: A Computational Perspective. *Acm Transactions On Intelligent Systems And Technology* (2021). <https://doi.org/10.1145/3546872>
- [296] Jiongnan Liu, Jiajie Jin, Zihan Wang, Jiehan Cheng, Zhicheng Dou, and Ji-Rong Wen. 2023. RETA-LLM: A Retrieval-Augmented Large Language Model Toolkit. *arXiv preprint arXiv: 2306.05212* (2023).
- [297] Q. Liu, Jun Wu, Shu Wu, and Liang Wang. 2023. Out-of-distribution Evidence-aware Fake News Detection via Dual Adversarial Debiasing. *ARXIV.ORG* (2023). <https://doi.org/10.48550/arXiv.2304.12888>
- [298] Xiao Liu, Hanyu Lai, Hao Yu, Yifan Xu, Aohan Zeng, Zhengxiao Du, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. WebGLM: Towards An Efficient Web-Enhanced Question Answering System with Human Preferences. *arXiv preprint arXiv: 2306.07906* (2023).
- [299] Xin Liu, Daniel McDuff, Geza Kovacs, Isaac Galatzer-Levy, Jacob Sunshine, Jiening Zhan, Ming-Zher Poh, Shun Liao, Paolo Di Achille, and Shwetak Patel. 2023. Large Language Models are Few-Shot Health Learners. *arXiv preprint arXiv: 2305.15525* (2023).
- [300] Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2023. AutoDAN: Generating Stealthy Jailbreak Prompts on Aligned Large Language Models. *arXiv preprint arXiv: 2310.04451* (2023).
- [301] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. 2023. AgentBench: Evaluating LLMs as Agents. *arXiv preprint arXiv: 2308.03688* (2023).
- [302] Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, and Yang Liu. 2023. Prompt Injection attack against LLM-integrated Applications. *arXiv preprint arXiv: 2306.05499* (2023).
- [303] Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, and Yang Liu. 2023. Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study. *arXiv preprint arXiv: 2305.13860* (2023).
- [304] Yang Liu and Yi-Fang Brook Wu. 2020. FNED: A Deep Network for Fake News Early Detection on Social Media. *ACM Trans. Inf. Syst.* 38, 3, Article 25 (2020), 33 pages. <https://doi.org/10.1145/3386253>
- [305] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models’ Alignment. *arXiv preprint arXiv: 2308.05374* (2023).
- [306] Ye Liu, Semih Yavuz, Rui Meng, Meghana Moorthy, Shafiq Joty, Caiming Xiong, and Yingbo Zhou. 2023. Exploring the Integration Strategies of Retriever and Large Language Models. *arXiv preprint arXiv: 2308.12574* (2023).
- [307] Yikang Liu, Ziyin Zhang, Wanyang Zhang, Shisen Yue, Xiaojing Zhao, Xinyuan Cheng, Yiwen Zhang, and Hai Hu. 2023. ArguGPT: evaluating, understanding and identifying argumentative essays generated by GPT models. *ARXIV.ORG* (2023). <https://doi.org/10.48550/arXiv.2304.07666>
- [308] Zhaoyang Liu, Yinan He, Wenhai Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Zeqiang Lai, Yang Yang, Qingyun Li, Jiashuo Yu, Kunchang Li, Zhe Chen, Xue Yang, Xizhou Zhu, Yali Wang, Limin Wang, Ping Luo, Jifeng Dai, and Yu Qiao. 2023. InternGPT: Solving Vision-Centric Tasks by Interacting with ChatGPT Beyond Language. *arXiv preprint arXiv: 2305.05662* (2023).
- [309] Zhiwei Liu, Weiran Yao, Jianguo Zhang, Le Xue, Shelby Heinecke, Rithesh Murthy, Yihao Feng, Zeyuan Chen, Juan Carlos Niebles, Devansh Arpit, Ran Xu, Phil Mui, Huan Wang, Caiming Xiong, and Silvio Savarese. 2023. BOLAA: Benchmarking and Orchestrating LLM-augmented Autonomous Agents. *arXiv preprint arXiv: 2308.05960* (2023).
- [310] Sahil Loomba, Alexandre de Figueiredo, Simon J Piatek, Kristen de Graaf, and Heidi J Larson. 2021. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature human behaviour* 5, 3 (2021), 337–348.
- [311] Menglong Lu, Zhen Huang, Binyang Li, Yunxiang Zhao, Zheng Qin, and Dongsheng Li. 2022. SIFTER: A Framework for Robust Rumor Detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), 429–442. <https://doi.org/10.1109/TASLP.2022.3140474>
- [312] Qinghua Lu, Liming Zhu, Xiwei Xu, Jon Whittle, Didar Zowghi, and Aurelie Jacquet. 2022. Responsible AI Pattern Catalogue: A Collection of Best Practices for AI Governance and Engineering. *arXiv preprint arXiv: 2209.04963* (2022).
- [313] Yi-Ju Lu and Cheng-Te Li. 2020. GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 505–514. <https://doi.org/10.18653/v1/2020.acl-main.48>
- [314] Guoqing Luo, Yu Tong Han, Lili Mou, and Mauajama Firdaus. 2023. Prompt-Based Editing for Text Style Transfer. *arXiv preprint arXiv: 2301.11997* (2023).
- [315] Junyu Luo, Cao Xiao, and Fenglong Ma. 2023. Zero-Resource Hallucination Prevention for Large Language Models. *arXiv preprint arXiv: 2309.02654* (2023).
- [316] Yuefei Lyu, Xiaoyu Yang, Jiaxin Liu, Sihong Xie, Philip Yu, and Xi Zhang. 2023. Interpretable and effective reinforcement learning for attacking against graph-based rumor detection. In *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–9.
- [317] Jing Ma and Wei Gao. 2020. Debunking Rumors on Twitter with Tree Transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 5455–5466. <https://doi.org/10.18653/v1/2020.coling-main.476>
- [318] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting Rumors from Microblogs with Recurrent Neural Networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, Subbarao Kambhampati (Ed.). IJCAI/AAAI Press, 3818–3824. <http://www.ijcai.org/Abstract/16/537>
- [319] Jing Ma, Jun Li, Wei Gao, Yang Yang, and Kam-Fai Wong. 2021. Improving rumor detection by promoting information campaigns with transformer-based generative adversarial learning. *IEEE Transactions*

- on *Knowledge and Data Engineering* (2021).
- [320] Jiachen Ma, Yong Liu, Meng Liu, and Meng Han. 2022. Curriculum Contrastive Learning for Fake News Detection. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (Atlanta, GA, USA) (CIKM '22). Association for Computing Machinery, New York, NY, USA, 4309–4313. <https://doi.org/10.1145/3511808.3557574>
  - [321] Yongqiang Ma, Jiawei Liu, Fan Yi, Qikai Cheng, Yong Huang, Wei Lu, and Xiaozhong Liu. 2023. AI vs. Human - Differentiation Analysis of Scientific Content Generation. *arXiv preprint arXiv: Arxiv-2301.10416* (2023).
  - [322] Abdurahman Maarouf, Dominik Bär, Dominique Geissler, and Stefan Feuerriegel. 2023. HQP: A Human-Annotated Dataset for Detecting Online Propaganda. *arXiv preprint arXiv: 2304.14931* (2023).
  - [323] Syed Mahbub, Eric Pardede, and ASM Kayes. 2022. COVID-19 Rumor Detection Using Psycho-Linguistic Features. *IEEE Access* 10 (2022), 117530–117543.
  - [324] Mohammad Mahyoub, Jeehaan Al-Garaady, and Musaad Alrahaile. 2020. Linguistic-based detection of fake news in social media. *Forthcoming, International Journal of English Linguistics* 11, 1 (2020).
  - [325] Martin Májovský, Martin Černý, Matěj Kasal, Martin Komarc, and David Netuka. 2023. Artificial Intelligence Can Generate Fraudulent but Authentic-Looking Scientific Medical Articles: Pandora’s Box Has Been Opened. *Journal of Medical Internet Research* 25 (2023), e46924.
  - [326] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Self-CheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. *arXiv preprint arXiv: Arxiv-2303.08896* (2023).
  - [327] Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020. A Survey on Computational Propaganda Detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, Christian Bessiere (Ed.). ijcai.org, 4826–4832. <https://doi.org/10.24963/ijcai.2020/672>
  - [328] Hana Matatov, Mor Naaman, and Ofra Amir. 2022. Stop the [Image] Steal: The Role and Dynamics of Visual Content in the 2020 U.S. Election Misinformation Campaign. *arXiv preprint arXiv: 2209.02007* (2022).
  - [329] Mohit Mayank, Shakshi Sharma, and Rajesh Sharma. 2021. DEAP-FAKED: Knowledge Graph based Approach for Fake News Detection. *International Conference on Advances in Social Networks Analysis and Mining* (2021). <https://doi.org/10.1109/ASONAM55673.2022.10068653>
  - [330] Philip M McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph. D. Dissertation. The University of Memphis.
  - [331] 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 (2020), 112986.
  - [332] Ninareh Mehrabi, Palash Goyal, Christophe Dupuy, Qian Hu, Shalini Ghosh, Richard Zemel, Kai-Wei Chang, Aram Galstyan, and Rahul Gupta. 2023. FLIRT: Feedback Loop In-context Red Teaming. *arXiv preprint arXiv: 2308.04265* (2023).
  - [333] Nikhil Mehta, Maria Leonor Pacheco, and Dan Goldwasser. 2022. Tackling Fake News Detection by Continually Improving Social Context Representations using Graph Neural Networks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 1363–1380. <https://doi.org/10.18653/v1/2022.acl-long.97>
  - [334] Alex Mei, Sharon Levy, and William Yang Wang. 2023. ASSERT: Automated Safety Scenario Red Teaming for Evaluating the Robustness of Large Language Models. *arXiv preprint arXiv: 2310.09624* (2023).
  - [335] Filippo Menczer, David Crandall, Yong-Yeol Ahn, and Apu Kapadia. 2023. Addressing the harms of AI-generated inauthentic content. *Nature Machine Intelligence* 5, 7 (2023), 679–680.
  - [336] Ethan Mendes, Yang Chen, Wei Xu, and Alan Ritter. 2023. Human-in-the-loop Evaluation for Early Misinformation Detection: A Case Study of COVID-19 Treatments. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 15817–15835. <https://doi.org/10.18653/v1/2023.acl-long.881>
  - [337] Kevin Meng, David Bau, A. Andonian, and Yonatan Belinkov. 2022. Locating and Editing Factual Associations in GPT. *Neural Information Processing Systems* (2022).
  - [338] Grégoire Mialon, Roberto Dessi, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023. Augmented Language Models: a Survey. *arXiv preprint arXiv: 2302.07842* (2023).
  - [339] midjourney. 2023. midjourney.com. <https://www.midjourney.com/>. Accessed: 2023-10-3.
  - [340] Erxue Min, Yu Rong, Yatao Bian, Tingyang Xu, Peilin Zhao, Junzhou Huang, and Sophia Ananiadou. 2022. Divide-and-Conquer: Post-User Interaction Network for Fake News Detection on Social Media. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) (WWW '22). Association for Computing Machinery, New York, NY, USA, 1148–1158. <https://doi.org/10.1145/3485447.3512163>
  - [341] Fatemehsadat Mireshghallah, Justus Mattern, Sicun Gao, Reza Shokri, and Taylor Berg-Kirkpatrick. 2023. Smaller Language Models are Better Black-box Machine-Generated Text Detectors. *arXiv preprint arXiv: 2305.09859* (2023).
  - [342] Saurabh Mishra, Jack Clark, and C. Raymond Perrault. 2020. Measurement in AI Policy: Opportunities and Challenges. *arXiv preprint arXiv: 2009.09071* (2020).
  - [343] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. *arXiv preprint arXiv: Arxiv-2301.11305* (2023).
  - [344] Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. 2023. ChatGPT or Human? Detect and Explain. Explaining Decisions of Machine Learning Model for Detecting Short ChatGPT-generated Text. *ARXIV.ORG* (2023). <https://doi.org/10.48550/arXiv.2301.13852>
  - [345] Ryan C Moore, Ross Dahlke, and Jeffrey T Hancock. 2023. Exposure to untrustworthy websites in the 2020 US election. *Nature Human Behaviour* (2023), 1–10.
  - [346] Ahmadreza Mosallanezhad, Mansoor Karami, Kai Shu, Michelle V Mancenido, and Huan Liu. 2022. Domain adaptive fake news detection via reinforcement learning. In *Proceedings of the ACM Web Conference 2022*. 3632–3640.
  - [347] Maximilian Mozes, Xuanli He, Bennett Kleinberg, and Lewis D. Griffin. 2023. Use of LLMs for Illicit Purposes: Threats, Prevention Measures, and Vulnerabilities. *arXiv preprint arXiv: 2308.12833* (2023).
  - [348] Akhtar Mubashara, Schlichtkrull Michael, Guo Zhiqiang, Cocarascu Oana, Simperl Elena, and Vlachos Andreas. 2023. Multimodal Automated Fact-Checking: A Survey. *arXiv preprint arXiv: 2305.13507* (2023).
  - [349] Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2023. Generating Benchmarks for Factuality Evaluation of Language Models. *arXiv preprint arXiv: 2307.06908* (2023).
  - [350] Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2023. Contrasting Linguistic Patterns in Human and LLM-Generated Text. *arXiv preprint arXiv: 2308.09067* (2023).
  - [351] Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory Hallucinations of Large Language Models: Evaluation, Detection and Mitigation. *arXiv preprint arXiv: 2305.15852*



- (2023).
- [352] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv: 2112.09332* (2021).
  - [353] Preslav Nakov, Alberto Barrón-Cedeño, Giovanni Da San Martino, Firoj Alam, Julia Maria Struß, Thomas Mandl, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouani, Chengkai Li, Shaden Shaar, Gautam Kishore Shahi, Hamdy Mubarak, Alex Nikolov, Nikolay Babulkov, Yavuz Selim Kartal, and Javier Beltrán. 2022. The CLEF-2022 CheckThat! Lab on Fighting the COVID-19 Infodemic and Fake News Detection. In *Advances in Information Retrieval*, Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty (Eds.). Springer International Publishing, Cham, 416–428.
  - [354] Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated Fact-Checking for Assisting Human Fact-Checkers. *arXiv preprint arXiv: 2103.07769* (2021).
  - [355] Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. MDFEND: Multi-Domain Fake News Detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (Virtual Event, Queensland, Australia) (CIKM '21). Association for Computing Machinery, New York, NY, USA, 3343–3347. <https://doi.org/10.1145/3459637.3482139>
  - [356] Christof Naumzik and Stefan Feuerriegel. 2022. Detecting False Rumors from Retweet Dynamics on Social Media. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) (WWW '22). Association for Computing Machinery, New York, NY, USA, 2798–2809. <https://doi.org/10.1145/3485447.3512000>
  - [357] Subhash Nerella, Sabyasachi Bandyopadhyay, Jiaqing Zhang, Miguel Contreras, Scott Siegel, Aysegül Bumin, Brandon Silva, Jessica Sena, Benjamin Shickel, Azra Bihorac, Kia Khezeli, and Parisa Rashidi. 2023. Transformers in Healthcare: A Survey. *arXiv preprint arXiv: 2307.00067* (2023).
  - [358] newsguardtech.com. 2023. Rise of the Newsbots: AI-Generated News Websites Proliferating Online. <https://www.newsguardtech.com/special-reports/newsbots-ai-generated-news-websites-proliferating/>. Accessed: 2023-10-3.
  - [359] An T Nguyen, Aditya Kharosekar, Saumya Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C Wallace, and Matthew Lease. 2018. Believe it or not: Designing a human-ai partnership for mixed-initiative fact-checking. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. 189–199.
  - [360] Duc Minh Nguyen, Tien Huu Do, Robert Calderbank, and Nikos Deligiannis. 2019. Fake News Detection using Deep Markov Random Fields. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 1391–1400. <https://doi.org/10.18653/v1/N19-1141>
  - [361] Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. FANG: Leveraging Social Context for Fake News Detection Using Graph Representation. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux (Eds.). ACM, 1165–1174. <https://doi.org/10.1145/3340531.3412046>
  - [362] Dan Saattrup Nielsen and Ryan McConville. 2022. MuMiN: A Large-Scale Multilingual Multimodal Fact-Checked Misinformation Social Network Dataset. *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (2022). <https://doi.org/10.1145/3477495.3531744>
  - [363] Ray Oshikawa, Jing Qian, and William Yang Wang. 2020. A Survey on Natural Language Processing for Fake News Detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 6086–6093. <https://aclanthology.org/2020.lrec-1.747>
  - [364] Nedjma Ousidhoum, Zhangdie Yuan, and Andreas Vlachos. 2022. Varifocal Question Generation for Fact-checking. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2532–2544. <https://aclanthology.org/2022.emnlp-main.163>
  - [365] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). <https://openreview.net/forum?id=TG8KACxEON>
  - [366] Leena Paakkari and Orkan Okan. 2020. COVID-19: health literacy is an underestimated problem. *The Lancet Public Health* 5, 5 (2020), e249–e250.
  - [367] Artidoro Pagnoni, Martin Graciarena, and Yulia Tsvetkov. 2022. Threat Scenarios and Best Practices to Detect Neural Fake News. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 1233–1249. <https://aclanthology.org/2022.coling-1.106>
  - [368] Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, Russa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, and Damien Graux. 2023. Large Language Models and Knowledge Graphs: Opportunities and Challenges. *arXiv preprint arXiv: 2308.06374* (2023).
  - [369] Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-Checking Complex Claims with Program-Guided Reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 6981–7004. <https://doi.org/10.18653/v1/2023.acl-long.386>
  - [370] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2023. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *arXiv preprint arXiv: 2306.08302* (2023).
  - [371] Wenjing Pan, Diyi Liu, and Jie Fang. 2021. An examination of factors contributing to the acceptance of online health misinformation. *Frontiers in psychology* 12 (2021), 630268.
  - [372] Yikang Pan, Liangming Pan, Wenhui 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* (2023).
  - [373] Subhadarshi Panda and Sarah Ita Levitan. 2021. Detecting Multilingual COVID-19 Misinformation on Social Media via Contextualized Embeddings. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*. Association for Computational Linguistics, Online, 125–129. <https://doi.org/10.18653/v1/2021.nlp4if-1.19>
  - [374] E. Papadogiannakis, P. Papadopoulos, Evangelos P. Markatos, and N. Kourtellis. 2022. Who Funds Misinformation? A Systematic Analysis of the Ad-related Profit Routines of Fake News sites. *The Web Conference* (2022). <https://doi.org/10.1145/3543507.3583443>

- [375] Shivam B Parikh and Pradeep K Atrey. 2018. Media-rich fake news detection: A survey. In *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*. IEEE, 436–441.
- [376] Peter S. Park, Simon Goldstein, Aidan O’Gara, Michael Chen, and Dan Hendrycks. 2023. AI Deception: A Survey of Examples, Risks, and Potential Solutions. *arXiv preprint arXiv: 2308.14752* (2023).
- [377] Ajay Patel, Nicholas Andrews, and Chris Callison-Burch. 2022. Low-Resource Authorship Style Transfer: Can Non-Famous Authors Be Imitated? *arXiv preprint arXiv: 2212.08986* (2022).
- [378] Ajay Patel, Delip Rao, and Chris Callison-Burch. 2023. Learning Interpretable Style Embeddings via Prompting LLMs. *arXiv preprint arXiv: 2305.12696* (2023).
- [379] Ajeet Ram Pathak, Aditee Mahajan, Keshav Singh, Aishwarya Patil, and Anusha Nair. 2020. Analysis of techniques for rumor detection in social media. *Procedia Computer Science* 167 (2020), 2286–2296.
- [380] Parth Patwa, Shivam Sharma, Srinivas PYKL, Vineeth Guptha, Gitanjali Kumari, Md. Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2021. Fighting an Infodemic: COVID-19 Fake News Dataset. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation - First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers (Communications in Computer and Information Science, Vol. 1402)*, Tanmoy Chakraborty, Kai Shu, H. Russell Bernard, Huan Liu, and Md. Shad Akhtar (Eds.). Springer, 21–29. [https://doi.org/10.1007/978-3-030-73696-5\\_3](https://doi.org/10.1007/978-3-030-73696-5_3)
- [381] Bohdan M. Pavlyshenko. 2023. Analysis of Disinformation and Fake News Detection Using Fine-Tuned Large Language Model. *arXiv preprint arXiv: 2309.04704* (2023).
- [382] Jessica Paynter, Sarah Luskin-Saxby, Deb Keen, Kathryn Fordyce, Grace Frost, Christine Imms, Scott Miller, David Trembath, Madonna Tucker, and Ullrich Ecker. 2019. Evaluation of a template for countering misinformation—Real-world Autism treatment myth debunking. *PLoS one* 14, 1 (2019), e0210746.
- [383] Kellin Pelrine, Jacob Danovitch, and Reihaneh Rabbany. 2021. The Surprising Performance of Simple Baselines for Misinformation Detection. In *Proceedings of the Web Conference 2021 (Ljubljana, Slovenia) (WWW ’21)*. Association for Computing Machinery, New York, NY, USA, 3432–3441. <https://doi.org/10.1145/3442381.3450111>
- [384] Kellin Pelrine, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, and Reihaneh Rabbany. 2023. Towards Reliable Misinformation Mitigation: Generalization, Uncertainty, and GPT-4. *arXiv preprint arXiv: 2305.14928* (2023).
- [385] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red Teaming Language Models with Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 3419–3448. <https://aclanthology.org/2022.emnlp-main.225>
- [386] Ethan Perez, Sam Ringer, Kamilë Lukošūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, C. McKinnon, C. Olah, Daisong Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, G. Khundadze, John Kernion, J. Landis, Jamie Kerr, J. Mueller, Jeeyoon Hyun, J. Landau, Kamal Ndousse, L. Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, S. Kravec, S. E. Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom B. Brown, T. Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Sam Bowman, Amanda Askell, Roger C. Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2022. Discovering Language Model Behaviors with Model-Written Evaluations. *Annual Meeting of the Association for Computational Linguistics* (2022). <https://doi.org/10.48550/arXiv.2212.09251>
- [387] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic Detection of Fake News. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 3391–3401. <https://aclanthology.org/C18-1287>
- [388] Roy H Perlis, Kristin Lunz Trujillo, Jon Green, Alauna Safarpour, James N Druckman, Mauricio Santillana, Katherine Ognyanova, and David Lazer. 2023. Misinformation, Trust, and Use of Ivermectin and Hydroxychloroquine for COVID-19. In *JAMA Health Forum*, Vol. 4. American Medical Association, e233257–e233257.
- [389] Heinrich Peters and Sandra Matz. 2023. Large Language Models Can Infer Psychological Dispositions of Social Media Users. *arXiv preprint arXiv: 2309.08631* (2023).
- [390] Francesco Pierri, Geng Liu, and Stefano Ceri. 2023. ITA-ELECTION-2022: A multi-platform dataset of social media conversations around the 2022 Italian general election. *arXiv preprint arXiv: 2301.05119* (2023).
- [391] Francesco Pierri, Luca Luceri, Nikhil Jindal, and Emilio Ferrara. 2022. Propaganda and Misinformation on Facebook and Twitter during the Russian Invasion of Ukraine. *Web Science Conference* (2022). <https://doi.org/10.1145/3578503.3583597>
- [392] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 22–32. <https://doi.org/10.18653/v1/D18-1003>
- [393] Piotr Przybyla. 2020. Capturing the Style of Fake News. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 490–497. <https://aaai.org/ojs/index.php/AAAI/article/view/5386>
- [394] Peng Qi, Juan Cao, Xirong Li, Huan Liu, Qiang Sheng, Xiaoyue Mi, Qin He, Yongbiao Lv, Chenyang Guo, and Yingchao Yu. 2021. Improving Fake News Detection by Using an Entity-Enhanced Framework to Fuse Diverse Multimodal Clues. In *Proceedings of the 29th ACM International Conference on Multimedia (Virtual Event, China) (MM ’21)*. Association for Computing Machinery, New York, NY, USA, 1212–1220. <https://doi.org/10.1145/3474085.3481548>
- [395] Peng Qi, Yuyang Zhao, Yufeng Shen, Wei Ji, Juan Cao, and Tat-Seng Chua. 2023. Two Heads Are Better Than One: Improving Fake News Video Detection by Correlating with Neighbors. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, Toronto, Canada, 11947–11959. <https://doi.org/10.18653/v1/2023.findings-acl.756>
- [396] Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Mengdi Wang, and Prateek Mittal. 2023. Visual Adversarial Examples Jailbreak Large Language Models. *arXiv preprint arXiv: 2306.13213* (2023).
- [397] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! *arXiv preprint arXiv: 2310.03693* (2023).
- [398] Cheng Qian, Chi Han, Yi R. Fung, Yujia Qin, Zhiyuan Liu, and Heng Ji. 2023. CREATOR: Disentangling Abstract and Concrete Reasonings of Large Language Models through Tool Creation. *arXiv preprint arXiv: 2305.14318* (2023).
- [399] Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical Multi-Modal Contextual Attention

- Network for Fake News Detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) (SIGIR '21). Association for Computing Machinery, New York, NY, USA, 153–162. <https://doi.org/10.1145/3404835.3462871>
- [400] Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with Language Model Prompting: A Survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 5368–5393. <https://doi.org/10.18653/v1/2023.acl-long.294>
- [401] Yu Qiao, Daniel Wiechmann, and Elma Kerz. 2020. A Language-Based Approach to Fake News Detection Through Interpretable Features and BRNN. In *Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM)*. Association for Computational Linguistics, Barcelona, Spain (Online), 14–31. <https://aclanthology.org/2020.rdsm-1.2>
- [402] Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, Y. Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shi Liang, Xingyu Shen, Bokai Xu, Zhen Zhang, Yining Ye, Bo Li, Ziwei Tang, Jing Yi, Yu Zhu, Zhenning Dai, Lan Yan, Xin Cong, Ya-Ting Lu, Weilin Zhao, Yuxiang Huang, Jun-Han Yan, Xu Han, Xian Sun, Dahai Li, Jason Phang, Cheng Yang, Tongshuang Wu, Heng Ji, Zhiyuan Liu, and Maosong Sun. 2023. Tool Learning with Foundation Models. *ARXIV.ORG* (2023). <https://doi.org/10.48550/arXiv.2304.08354>
- [403] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs. *arXiv preprint arXiv: 2307.16789* (2023).
- [404] Huachuan Qiu, Shuai Zhang, Anqi Li, Hongliang He, and Zhenzhong Lan. 2023. Latent Jailbreak: A Benchmark for Evaluating Text Safety and Output Robustness of Large Language Models. *arXiv preprint arXiv: 2307.08487* (2023).
- [405] Dorian Quelle and Alexandre Bovet. 2023. The Perils & Promises of Fact-checking with Large Language Models. *arXiv preprint arXiv:2310.13549* (2023).
- [406] P. Ranade, Aritrano Piplai, Sudip Mittal, A. Joshi, and Tim Finin. 2021. Generating Fake Cyber Threat Intelligence Using Transformer-Based Models. *IEEE International Joint Conference on Neural Network* (2021). <https://doi.org/10.1109/IJCNN52387.2021.9534192>
- [407] Aman Rangapur, Haoran Wang, and Kai Shu. 2023. Investigating Online Financial Misinformation and Its Consequences: A Computational Perspective. *arXiv preprint arXiv: 2309.12363* (2023).
- [408] Abhinav Rao, Sachin Vashistha, Atharva Naik, Somak Aditya, and Monojit Choudhury. 2023. Tricking LLMs into Disobedience: Understanding, Analyzing, and Preventing Jailbreaks. *arXiv preprint arXiv: 2305.14965* (2023).
- [409] Dongning Rao, Xin Miao, Zhihua Jiang, and Ran Li. 2021. STANKER: Stacking Network based on Level-grained Attention-masked BERT for Rumor Detection on Social Media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 3347–3363. <https://doi.org/10.18653/v1/2021.emnlp-main.269>
- [410] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 2931–2937. <https://doi.org/10.18653/v1/D17-1317>
- [411] Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, Vol. 11. NIH Public Access, 269.
- [412] Maribeth Rauh, John Mellor, Jonathan Uesato, Po-Sen Huang, Johannes Welbl, Laura Weidinger, Sumanth Dathathri, Amelia Glaese, Geoffrey Irving, Iason Gabriel, William Isaac, and Lisa Anne Hendricks. 2022. Characteristics of Harmful Text: Towards Rigorous Benchmarking of Language Models. In *NeurIPS*. [http://papers.nips.cc/paper\\_files/paper/2022/hash/9ca22870ae0ba55ee50ce3e2d269e5de-Abstract-Datasets\\_and\\_Benchmarks.html](http://papers.nips.cc/paper_files/paper/2022/hash/9ca22870ae0ba55ee50ce3e2d269e5de-Abstract-Datasets_and_Benchmarks.html)
- [413] Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S. M Towhidul Islam Tonmoy, Aman Chadha, Amit P. Sheth, and Amitava Das. 2023. The Troubling Emergence of Hallucination in Large Language Models - An Extensive Definition, Quantification, and Prescriptive Remediations. *arXiv preprint arXiv: 2310.04988* (2023).
- [414] Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A Survey of Hallucination in Large Foundation Models. *arXiv preprint arXiv: 2309.05922* (2023).
- [415] Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A Recipe for Arbitrary Text Style Transfer with Large Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Dublin, Ireland, 837–848. <https://doi.org/10.18653/v1/2022.acl-short.94>
- [416] Yuxiang Ren, Bo Wang, Jiawei Zhang, and Yi Chang. 2020. Adversarial Active Learning Based Heterogeneous Graph Neural Network for Fake News Detection. In *2020 IEEE International Conference on Data Mining (ICDM)*. 452–461. <https://doi.org/10.1109/ICDM50108.2020.00054>
- [417] Alexander Robey, Eric Wong, Hamed Hassani, and George J. Pappas. 2023. SmoothLLM: Defending Large Language Models Against Jailbreaking Attacks. *arXiv preprint arXiv: 2310.03684* (2023).
- [418] Natalia Diaz Rodriguez, J. Ser, M. Coeckelbergh, Marcos L’opez de Prado, E. Herrera-Viedma, and Francisco Herrera. 2023. Connecting the Dots in Trustworthy Artificial Intelligence: From AI Principles, Ethics, and Key Requirements to Responsible AI Systems and Regulation. *ARXIV.ORG* (2023). <https://doi.org/10.48550/arXiv.2305.02231>
- [419] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and B. Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *Computer Vision and Pattern Recognition* (2021). <https://doi.org/10.1109/CVPR52688.2022.01042>
- [420] Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*. Association for Computational Linguistics, San Diego, California, 7–17. <https://doi.org/10.18653/v1/W16-0802>
- [421] Arkadiy Saakyan and Smaranda Muresan. 2023. ICLEF: In-Context Learning with Expert Feedback for Explainable Style Transfer. *arXiv preprint arXiv: 2309.08583* (2023).
- [422] Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can AI-Generated Text be Reliably Detected? *arXiv preprint arXiv: Arxiv-2303.11156* (2023).
- [423] Ujala Sajid and Faheem ul Hassan. 2022. ChatGPT and its effect on Shaping the Future of Medical Writing. *Pakistan Journal of Ethics* 2, 2 (2022), 38–43.
- [424] Mansi Sakarvadia, Aswathy Ajith, Arham Khan, Daniel Grzenda, Nathaniel Hudson, André Bauer, Kyle Chard, and Ian Foster. 2023. Memory Injections: Correcting Multi-Hop Reasoning Failures during Inference in Transformer-Based Language Models. *arXiv preprint arXiv: 2309.05605* (2023).

- [425] Nikos Salamanos, Pantelitsa Leonidou, Nikolaos Laoutaris, Michael Sirivianos, Maria Aspri, and Marius Paraschiv. 2023. HyperGraphDis: Leveraging Hypergraphs for Contextual and Social-Based Disinformation Detection. *arXiv preprint arXiv: 2310.01113* (2023).
- [426] Malik Sallam. 2023. The Utility of ChatGPT as an Example of Large Language Models in Healthcare Education, Research and Practice: Systematic Review on the Future Perspectives and Potential Limitations. *medRxiv* (2023), 2023–02.
- [427] Emily Saltz, Soubhik Barari, Claire Leibowicz, and Claire Wardle. 2021. Misinformation interventions are common, divisive, and poorly understood. *Harvard Kennedy School (HKS) Misinformation Review* 2, 5 (2021).
- [428] Emily Saltz, Claire R Leibowicz, and Claire Wardle. 2021. Encounters with visual misinformation and labels across platforms: An interview and diary study to inform ecosystem approaches to misinformation interventions. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–6.
- [429] Mourad Sarroui, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based Fact-Checking of Health-related Claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 3499–3512. <https://doi.org/10.18653/v1/2021.findings-emnlp.297>
- [430] Dietram A Scheufele, Andrew J Hoffman, Liz Neeley, and Czerne M Reid. 2021. Misinformation about science in the public sphere. *Proceedings of the National Academy of Sciences* 118, 15 (2021), e2104068118.
- [431] Jonas Schuett. 2019. Defining the scope of AI regulations. *arXiv preprint arXiv: 1909.01095* (2019).
- [432] Tal Schuster, Roei Schuster, Darsh J. Shah, and Regina Barzilay. 2020. The Limitations of Stylometry for Detecting Machine-Generated Fake News. *Computational Linguistics* 46, 2 (2020), 499–510. [https://doi.org/10.1162/coli\\_a\\_00380](https://doi.org/10.1162/coli_a_00380)
- [433] Stephane Schwarz, Antônio Theóphilo, and Anderson Rocha. 2020. EMET: Embeddings from Multilingual-Encoder Transformer for Fake News Detection. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*. IEEE, 2777–2781. <https://doi.org/10.1109/ICASSP40776.2020.9054673>
- [434] scmp.com. 2023. Chinese artificial intelligence firm iFlytek blames chatbot-generated article for sudden share price swing on Shenzhen bourse. <https://www.scmp.com/tech/big-tech/article/3221953/chinese-artificial-intelligence-firm-iflytek-blames-chatbot-generated-article-sudden-share-price>. Accessed: 2023-09-30.
- [435] Elizabeth Seger, Noemi Dreksler, Richard Moulange, Emily Dardaman, Jonas Schuett, K Wei, Christoph Winter, Mackenzie Arnold, Seán Ó hÉigeartaigh, Anton Korinek, et al. 2023. Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives. (2023).
- [436] Vinay Setty and Erlend Rekve. 2020. Truth be Told: Fake News Detection Using User Reactions on Reddit. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux (Eds.). ACM, 3325–3328. <https://doi.org/10.1145/3340531.3417463>
- [437] Shaden Shaar, Firoj Alam, Giovanni Da San Martino, and Preslav Nakov. 2022. The Role of Context in Detecting Previously Fact-Checked Claims. In *Findings of the Association for Computational Linguistics: NAACL 2022*. Association for Computational Linguistics, Seattle, United States, 1619–1631. <https://doi.org/10.18653/v1/2022.findings-naacl.122>
- [438] Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. 2022. A Duo-Generative Approach to Explainable Multimodal COVID-19 Misinformation Detection. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) (WWW '22). Association for Computing Machinery, New York, NY, USA, 3623–3631. <https://doi.org/10.1145/3485447.3512257>
- [439] Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy. *arXiv preprint arXiv: 2305.15294* (2023).
- [440] 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 (2019), 1–42.
- [441] Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023. Large Language Model Alignment: A Survey. *arXiv preprint arXiv: 2309.15025* (2023).
- [442] Xudong Shen, Hannah Brown, Jiashu Tao, Martin Strobel, Yao Tong, Akshay Narayan, Harold Soh, and Finale Doshi-Velez. 2023. Towards Regulatable AI Systems: Technical Gaps and Policy Opportunities. *arXiv preprint arXiv: 2306.12609* (2023).
- [443] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2023. "Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models. *arXiv preprint arXiv: 2308.03825* (2023).
- [444] Qiang Sheng, Juan Cao, Xueyao Zhang, Rundong Li, Danding Wang, and Yongchun Zhu. 2022. Zoom Out and Observe: News Environment Perception for Fake News Detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 4543–4556. <https://doi.org/10.18653/v1/2022.acl-long.311>
- [445] Qiang Sheng, Xueyao Zhang, Juan Cao, and Lei Zhong. 2021. Integrating Pattern- and Fact-Based Fake News Detection via Model Preference Learning. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (Virtual Event, Queensland, Australia) (CIKM '21)*. Association for Computing Machinery, New York, NY, USA, 1640–1650. <https://doi.org/10.1145/3459637.3482440>
- [446] Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, Lewis Ho, Divya Siddarth, Shahar Avin, Will Hawkins, Been Kim, Iason Gabriel, Vijay Bolina, Jack Clark, Yoshua Bengio, Paul Christiano, and Allan Dafoe. 2023. Model evaluation for extreme risks. *arXiv preprint arXiv: 2305.15324* (2023).
- [447] Zhouxing Shi, Yihan Wang, Fan Yin, Xiangning Chen, Kai-Wei Chang, and Cho-Jui Hsieh. 2023. Red Teaming Language Model Detectors with Language Models. *arXiv preprint arXiv: 2305.19713* (2023).
- [448] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. dFEND: Explainable Fake News Detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, Ankur Teredesai, Vipin Kumar, Ying Li, Rómer Rosales, Evimaria Terzi, and George Karypis (Eds.). ACM, 395–405. <https://doi.org/10.1145/3292500.3330935>
- [449] Kai Shu, Susan Dumais, Ahmed Hassan Awadallah, and Huan Liu. 2020. Detecting fake news with weak social supervision. *IEEE Intelligent Systems* 36, 4 (2020), 96–103.
- [450] Kai Shu, Yichuan Li, Kaize Ding, and Huan Liu. 2021. Fact-Enhanced Synthetic News Generation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 13825–13833. <https://ojs.aaai.org/index.php/AAAI/article/view/17629>
- [451] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective.

ACM SIGKDD explorations newsletter 19, 1 (2017), 22–36.

- [452] Kai Shu, Suhang Wang, Thai Le, Dongwon Lee, and Huan Liu. 2018. Deep headline generation for clickbait detection. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 467–476.
- [453] Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond News Contents: The Role of Social Context for Fake News Detection. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11–15, 2019*, J. Shane Culpepper, Alistair Moffat, Paul N. Bennett, and Kristina Lerman (Eds.). ACM, 312–320. <https://doi.org/10.1145/3289600.3290994>
- [454] Kai Shu, Guoqing Zheng, Yichuan Li, Subhabrata Mukherjee, Ahmed Hassan Awadallah, Scott Ruston, and Huan Liu. 2021. Early detection of fake news with multi-source weak social supervision. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part III*. Springer, 650–666.
- [455] Manli Shu, Jiong Xiao Wang, Chen Zhu, Jonas Geiping, Chaowei Xiao, and Tom Goldstein. 2023. On the Exploitability of Instruction Tuning. *arXiv preprint arXiv: 2306.17194* (2023).
- [456] Elena Shushkevich, Mikhail Alexandrov, and John Cardiff. 2023. Improving Multiclass Classification of Fake News Using BERT-Based Models and ChatGPT-Augmented Data. *Inventions* 8, 5 (2023), 112.
- [457] Amila Silva, Ling Luo, Shanika Karunasekera, and Christopher Leckie. 2021. Embracing Domain Differences in Fake News: Cross-domain Fake News Detection using Multi-modal Data. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2–9, 2021*. AAAI Press, 557–565. <https://ojs.aaai.org/index.php/AAAI/article/view/16134>
- [458] Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Hal Daumé III, Jesse Dodge, Ellie Evans, Sara Hooker, Yacine Jernite, Alexandra Sasha Luccioni, Alberto Lusoli, Margaret Mitchell, Jessica Newman, Marie-Therese Png, Andrew Strait, and Apostol Vassilev. 2023. Evaluating the Social Impact of Generative AI Systems in Systems and Society. *arXiv preprint arXiv: 2306.05949* (2023).
- [459] Chenguang Song, Kai Shu, and Bin Wu. 2021. Temporally evolving graph neural network for fake news detection. *Information Processing & Management* 58, 6 (2021), 102712.
- [460] Changhe Song, Cheng Yang, Huimin Chen, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2021. CED: Credible Early Detection of Social Media Rumors. *IEEE Transactions on Knowledge and Data Engineering* 33, 8 (2021), 3035–3047. <https://doi.org/10.1109/TKDE.2019.2961675>
- [461] Yun-Zhu Song, Yi-Syuan Chen, Yi-Ting Chang, Shao-Yu Weng, and Hong-Han Shuai. 2021. Adversary-Aware Rumor Detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, Online, 1371–1382. <https://doi.org/10.18653/v1/2021.findings-acl.118>
- [462] Damiano Spina, Mark Sanderson, Daniel Angus, Gianluca Demartini, Dana McKay, Lauren L Saling, and Ryan W White. 2023. Human-AI Cooperation to Tackle Misinformation and Polarization. *Commun. ACM* 66, 7 (2023), 40–45.
- [463] Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2023. AI model GPT-3 (dis)informs us better than humans. *Science Advances* 9, 26 (2023), eadh1850. <https://doi.org/10.1126/sciadv.adh1850> <https://www.science.org/doi/pdf/10.1126/sciadv.adh1850>
- [464] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askill, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubakaran, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Froberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqi, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Śwędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimeo Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdah Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas



- Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Milkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel Stern Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Mishnerghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, vinay uday prabhu, Vishakh Padmakumar, Vivek Sriku-mar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research* (2023). <https://openreview.net/forum?id=uyTL5Bvosj>
- [465] Logan Stapleton, Jordan Taylor, Sarah Fox, Tongshuang Wu, and Haiyi Zhu. 2023. Seeing Seeds Beyond Weeds: Green Teaming Generative AI for Beneficial Uses. *arXiv preprint arXiv: 2306.03097* (2023).
- [466] Kate Starbird, Renée DiResta, and Matt DeButts. 2023. Influence and improvisation: Participatory disinformation during the 2020 US election. *Social Media+ Society* 9, 2 (2023), 20563051231177943.
- [467] Harald Stiff and Fredrik Johansson. 2022. Detecting computer-generated disinformation. *International Journal of Data Science and Analytics* 13, 4 (2022), 363–383.
- [468] Jinyan Su, Terry Yue Zhuo, Jonibek Mansurov, Di Wang, and Preslav Nakov. 2023. Fake News Detectors are Biased against Texts Generated by Large Language Models. *arXiv preprint arXiv: 2309.08674* (2023).
- [469] Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. 2023. DetectLLM: Leveraging Log Rank Information for Zero-Shot Detection of Machine-Generated Text. *arXiv preprint arXiv: 2306.05540* (2023).
- [470] Xing Su, Jian Yang, Jia Wu, and Zitai Qiu. 2023. Hy-DeFake: Hypergraph Neural Networks for Detecting Fake News in Online Social Networks. *arXiv preprint arXiv: 2309.02692* (2023).
- [471] Xing Su, Jian Yang, Jia Wu, and Yuchen Zhang. 2023. Mining User-aware Multi-Relations for Fake News Detection in Large Scale Online Social Networks. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 51–59.
- [472] Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine L. Hermann, Kerem Oktar, Klaus Greff, Martin N. Hebart, Nori Jacoby, Qiuyi Zhang, Raja Marjeh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas P. O’Connell, Thomas Unterthiner, Andrew K. Lampinen, Klaus-Robert Müller, Mariya Toneva, and Thomas L. Griffiths. 2023. Getting aligned on representational alignment. *arXiv preprint arXiv: 2310.13018* (2023).
- [473] Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2023. Head-to-Tail: How Knowledgeable are Large Language Models (LLM)? A.K.A. Will LLMs Replace Knowledge Graphs? *arXiv preprint arXiv: 2308.10168* (2023).
- [474] Mengzhu Sun, Xi Zhang, Jianqiang Ma, and Yazheng Liu. 2021. Inconsistency Matters: A Knowledge-guided Dual-inconsistency Network for Multi-modal Rumor Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 1412–1423. <https://doi.org/10.18653/v1/2021.findings-emnlp.122>
- [475] Mengzhu Sun, Xi Zhang, Jiaqi Zheng, and Guixiang Ma. 2022. DDGCN: Dual Dynamic Graph Convolutional Networks for Rumor Detection on Social Media. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelfth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 4611–4619. <https://ojs.aaai.org/index.php/AAAI/article/view/20385>
- [476] Tiening Sun, Zhong Qian, Sujun Dong, Peifeng Li, and Qiaoming Zhu. 2022. Rumor Media Detection on Social Media with Graph Adversarial Contrastive Learning. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) (WWW ’22). Association for Computing Machinery, New York, NY, USA, 2789–2797. <https://doi.org/10.1145/3485447.3511999>
- [477] Yanshen Sun, Jianfeng He, Shuo Lei, Limeng Cui, and Chang-Tien Lu. 2023. Med-MMHL: A Multi-Modal Dataset for Detecting Human- and LLM-Generated Misinformation in the Medical Domain. *arXiv preprint arXiv: 2306.08871* (2023).
- [478] S Suryavardan, Shreyash Mishra, Megha Chakraborty, Parth Patwa, Anku Rani, Aman Chadha, Aishwarya Reganti, Amitava Das, Amit Sheth, Manoj Chinnakotla, Asif Ekbal, and Srijan Kumar. 2023. Findings of Factify 2: Multimodal Fake News Detection. *arXiv preprint arXiv: 2307.10475* (2023).
- [479] S Suryavardan, Shreyash Mishra, Parth Patwa, Megha Chakraborty, Anku Rani, Aishwarya Reganti, Aman Chadha, Amitava Das, Amit Sheth, Manoj Chinnakotla, Asif Ekbal, and Srijan Kumar. 2023. Factify 2: A Multimodal Fake News and Satire News Dataset. *arXiv preprint arXiv: 2304.03897* (2023).
- [480] Briony Swire-Thompson, Joseph DeGutis, and David Lazer. 2020. Searching for the backfire effect: Measurement and design considerations. *Journal of applied research in memory and cognition* 9, 3 (2020), 286–299.
- [481] Reuben Tan, Bryan Plummer, and Kate Saenko. 2020. Detecting Cross-Modal Inconsistency to Defend Against Neural Fake News. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 2081–2106. <https://doi.org/10.18653/v1/2020.emnlp-main.163>
- [482] Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023. Evaluation of ChatGPT as a Question Answering System for Answering Complex Questions. *arXiv preprint arXiv: Arxiv-2303.07992* (2023).
- [483] Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2023. The Science of Detecting LLM-Generated Texts. (2023). anonymous preprint under review.

- [484] Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. 2023. Multilingual LLMs are Better Cross-lingual In-context Learners with Alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, 6292–6307. <https://aclanthology.org/2023.acl-long.346>
- [485] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.
- [486] Yuchuan Tian, Hanting Chen, Xutao Wang, Zheyuan Bai, Qinghua Zhang, Ruifeng Li, Chao Xu, and Yunhe Wang. 2023. Multiscale Positive-Unlabeled Detection of AI-Generated Texts. *arXiv preprint arXiv: 2305.18149* (2023).
- [487] Ehsan Toreini, Mhairi Aitken, Kovila P. L. Coopamootoo, Karen Elliott, Vladimiro Gonzalez Zelaya, Paolo Missier, Magdalene Ng, and Aad van Moorsel. 2020. Technologies for Trustworthy Machine Learning: A Survey in a Socio-Technical Context. *arXiv preprint arXiv: 2007.08911* (2020).
- [488] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xi-ang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv: 2307.09288* (2023).
- [489] Robert Trager, Ben Harack, Anka Reuel, Allison Carnegie, Lennart Heim, Lewis Ho, Sarah Kreps, Ranjit Lall, Owen Larter, Seán Ó hÉigeartaigh, Simon Staffell, and José Jaime Villalobos. 2023. International Governance of Civilian AI: A Jurisdictional Certification Approach. *arXiv preprint arXiv: 2308.15514* (2023).
- [490] Milena Trajanoska, Riste Stojanov, and Dimitar Trajanov. 2023. Enhancing Knowledge Graph Construction Using Large Language Models. *arXiv preprint arXiv: 2305.04676* (2023).
- [491] Isaac Triguero, Daniel Molina, Javier Poyatos, Javier Del Ser, and Francisco Herrera. 2023. General Purpose Artificial Intelligence Systems (GPAIS): Properties, Definition, Taxonomy, Open Challenges and Implications. *arXiv preprint arXiv: 2307.14283* (2023).
- [492] Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct Distillation of LM Alignment. *arXiv preprint arXiv: 2310.16944* (2023).
- [493] Jacob Tyo, Bhuwan Dhingra, and Z. Lipton. 2022. On the State of the Art in Authorship Attribution and Authorship Verification. *ARXIV.ORG* (2022). <https://doi.org/10.48550/arXiv.2209.06869>
- [494] Adaku Uchendu, Thai Le, and Dongwon Lee. 2023. Attribution and Obfuscation of Neural Text Authorship: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter* 25, 1 (2023), 1–18.
- [495] Adaku Uchendu, Jooyoung Lee, Hua Shen, Thai Le, Ting-Hao 'Kenneth' Huang, and Dongwon Lee. 2023. Understanding Individual and Team-based Human Factors in Detecting Deepfake Texts. *ARXIV.ORG* (2023). <https://doi.org/10.48550/arXiv.2304.01002>
- [496] AR Ullah, Anupam Das, Anik Das, Muhammad Ashad Kabir, and Kai Shu. 2021. A survey of covid-19 misinformation: Datasets, detection techniques and open issues. *ArXiv preprint abs/2110.00737* (2021). <https://arxiv.org/abs/2110.00737>
- [497] Rishabh Upadhyay, Gabriella Pasi, and Marco Viviani. 2023. Leveraging Socio-Contextual Information in BERT for Fake Health News Detection in Social Media. In *Proceedings of the 3rd International Workshop on Open Challenges in Online Social Networks (Rome, Italy) (OASIS '23)*. Association for Computing Machinery, New York, NY, USA, 38–46. <https://doi.org/10.1145/3599696.3612902>
- [498] Vaibhav Vaibhav, Raghuram Mandyam, and Eduard Hovy. 2019. Do Sentence Interactions Matter? Leveraging Sentence Level Representations for Fake News Classification. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*. Association for Computational Linguistics, Hong Kong, 134–139. <https://doi.org/10.18653/v1/D19-5316>
- [499] Sander van der Linden. 2022. Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature Medicine* 28, 3 (2022), 460–467.
- [500] Francielle Vargas, Fabricio Benevenuto, and Thiago Pardo. 2021. Toward Discourse-Aware Models for Multilingual Fake News Detection. In *Proceedings of the Student Research Workshop Associated with RANLP 2021*. INCOMA Ltd., Online, 210–218. <https://aclanthology.org/2021.ranlp-srw.29>
- [501] Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence Generation. *arXiv preprint arXiv: 2307.03987* (2023).
- [502] Christoforos Vasilatos, Manaar Alam, Talal Rahwan, Yasir Zaki, and Michail Maniatakis. 2023. HowkGPT: Investigating the Detection of ChatGPT-generated University Student Homework through Context-Aware Perplexity Analysis. *arXiv preprint arXiv: 2305.18226* (2023).
- [503] Nikhita Vedula and Srinivasan Parthasarathy. 2021. FACE-KEG: Fact Checking Explained Using Knowledge Graphs. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining (Virtual Event, Israel) (WSDM '21)*. Association for Computing Machinery, New York, NY, USA, 526–534. <https://doi.org/10.1145/3437963.3441828>
- [504] Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2023. Ghostbuster: Detecting Text Ghostwritten by Large Language Models. *arXiv preprint arXiv: 2305.15047* (2023).
- [505] Prashanth Vijayaraghavan and Soroush Vosoughi. 2022. TWEET-SPIN: Fine-grained Propaganda Detection in Social Media Using Multi-View Representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 3433–3448. <https://doi.org/10.18653/v1/2022.naacl-main.251>
- [506] Juraj Vladika and F. Matthes. 2023. Scientific Fact-Checking: A Survey of Resources and Approaches. *Annual Meeting of the Association for Computational Linguistics* (2023). <https://doi.org/10.48550/arXiv.2305.16859>
- [507] Nguyen Vo and Kyumin Lee. 2020. Where Are the Facts? Searching for Fact-checked Information to Alleviate the Spread of Fake News. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 7717–7731. <https://doi.org/10.18653/v1/2020.emnlp-main.621>
- [508] Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2023. FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation. *arXiv preprint arXiv: 2310.03214* (2023).

- [509] Nathan Walter, John J Brooks, Camille J Saucier, and Sapna Suresh. 2021. Evaluating the impact of attempts to correct health misinformation on social media: A meta-analysis. *Health Communication* 36, 13 (2021), 1776–1784.
- [510] Nathan Walter and Sheila T Murphy. 2018. How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs* 85, 3 (2018), 423–441.
- [511] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. 2023. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. *arXiv preprint arXiv: 2306.11698* (2023).
- [512] Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023. Survey on Factuality in Large Language Models: Knowledge, Retrieval and Domain-Specificity. *arXiv preprint arXiv: 2310.07521* (2023).
- [513] Gengyu Wang, Kate Harwood, Lawrence Chillrud, Amith Ananthram, Melanie Subbiah, and Kathleen McKeown. 2023. Check-COVID: Fact-Checking COVID-19 News Claims with Scientific Evidence. *arXiv preprint arXiv: 2305.18265* (2023).
- [514] Gengyu Wang, Kate Harwood, Lawrence Chillrud, Amith Ananthram, Melanie Subbiah, and Kathleen McKeown. 2023. Check-COVID: Fact-Checking COVID-19 News Claims with Scientific Evidence. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, Toronto, Canada, 14114–14127. <https://doi.org/10.18653/v1/2023.findings-acl.888>
- [515] Haoran Wang, Yingdong Dou, Canyu Chen, Lichao Sun, Philip S. Yu, and Kai Shu. 2023. Attacking Fake News Detectors via Manipulating News Social Engagement. *The Web Conference* (2023). <https://doi.org/10.1145/3543507.3583868>
- [516] Haoran Wang and Kai Shu. 2023. Explainable Claim Verification via Knowledge-Grounded Reasoning with Large Language Models. *arXiv preprint arXiv: 2310.05253* (2023).
- [517] Jiaming Wang, Zhihao Du, Qian Chen, Yunfei Chu, Zhifu Gao, Zerui Li, Kai Hu, Xiaohuan Zhou, Jin Xu, Ziyang Ma, Wen Wang, Siqu Zheng, Chang Zhou, Zhijie Yan, and Shiliang Zhang. 2023. LauraGPT: Listen, Attend, Understand, and Regenerate Audio with GPT. *arXiv preprint arXiv: 2310.04673* (2023).
- [518] Jia Wang, Min Gao, Yinqiu Huang, Kai Shu, and Hualing Yi. 2023. FinD: Fine-grained discrepancy-based fake news detection enhanced by event abstract generation. *Computer Speech & Language* 78 (2023), 101461.
- [519] Lei Wang, Chengbang Ma, Xueyang Feng, Zeyu Zhang, Hao ran Yang, Jingsen Zhang, Zhi-Yang Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji rong Wen. 2023. A Survey on Large Language Model based Autonomous Agents. *ArXiv preprint abs/2308.11432* (2023). <https://arxiv.org/abs/2308.11432>
- [520] 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* (2023).
- [521] Ruize Wang, Duyu Tang, Nan Duan, Wanjun Zhong, Zhongyu Wei, Xuanjing Huang, Daxin Jiang, and Ming Zhou. 2020. Leveraging Declarative Knowledge in Text and First-Order Logic for Fine-Grained Propaganda Detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 3895–3903. <https://doi.org/10.18653/v1/2020.emnlp-main.320>
- [522] Tao Wang, Yushu Zhang, Shuren Qi, Ruoyu Zhao, Zhihua Xia, and Jian Weng. 2023. Security and Privacy on Generative Data in AIGC: A Survey. *arXiv preprint arXiv: 2309.09435* (2023).
- [523] Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen tse Huang, Wenxiang Jiao, and Michael R. Lyu. 2023. All Languages Matter: On the Multilingual Safety of Large Language Models. *arXiv preprint arXiv: 2310.00905* (2023).
- [524] Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. 2023. Do-Not-Answer: A Dataset for Evaluating Safeguards in LLMs. *arXiv preprint arXiv: 2308.13387* (2023).
- [525] Yi Wang, Kunchang Li, Yizhuo Li, Yanan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. 2022. InternVideo: General Video Foundation Models via Generative and Discriminative Learning. *arXiv preprint arXiv: 2212.03191* (2022).
- [526] Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023. Self-Knowledge Guided Retrieval Augmentation for Large Language Models. *arXiv preprint arXiv: 2310.05002* (2023).
- [527] Yaqing Wang, Fenglong Ma, Haoyu Wang, Kishlay Jha, and Jing Gao. 2021. Multimodal Emergent Fake News Detection via Meta Neural Process Networks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (Virtual Event, Singapore) (KDD '21)*. Association for Computing Machinery, New York, NY, USA, 3708–3716. <https://doi.org/10.1145/3447548.3467153>
- [528] Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Alham Fikri Aji, and Preslav Nakov. 2023. M4: Multi-generator, Multi-domain, and Multi-lingual Black-Box Machine-Generated Text Detection. *arXiv preprint arXiv: 2305.14902* (2023).
- [529] Yuntao Wang, Yanghe Pan, Miao Yan, Zhou Su, and Tom H. Luan. 2023. A Survey on ChatGPT: AI-Generated Contents, Challenges, and Solutions. *IEEE Open Journal of the Computer Society* (2023), 1–20. <https://doi.org/10.1109/OJCS.2023.3300321>
- [530] Yaqing Wang, Weifeng Yang, Fenglong Ma, Jin Xu, Bin Zhong, Qiang Deng, and Jing Gao. 2020. Weak Supervision for Fake News Detection via Reinforcement Learning. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7–12, 2020*. AAAI Press, 516–523. <https://aaai.org/ojs/index.php/AAAI/article/view/5389>
- [531] Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning Large Language Models with Human: A Survey. *arXiv preprint arXiv: 2307.12966* (2023).
- [532] Zecong Wang, Jiaxi Cheng, Chen Cui, and Chenhao Yu. 2023. Implementing BERT and fine-tuned RobertA to detect AI generated news by ChatGPT. *arXiv preprint arXiv: 2306.07401* (2023).
- [533] Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhang Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhui Chen, Jie Fu, and Junran Peng. 2023. RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models. *arXiv preprint arXiv: 2310.00746* (2023).
- [534] Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington. 2023. Testing of Detection Tools for AI-Generated Text. *arXiv preprint arXiv: 2306.15666* (2023).
- [535] Krzysztof Węcel, Marcin Sawiński, Milena Stróżyna, Włodzimierz Lewoniewski, Piotr Stolarski, Ewelina Książniak, and Witold Abramowicz. 2023. Artificial intelligence-friend or foe in fake news campaigns. *Economics and Business Review* 9, 2 (2023), 41–70.
- [536] Lingwei Wei, Dou Hu, Yantong Lai, Wei Zhou, and Songlin Hu. 2022. A Unified Propagation Forest-based Framework for Fake News Detection. In *Proceedings of the 29th International Conference on*

- Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2769–2779. <https://aclanthology.org/2022.coling-1.244>
- [537] Lingwei Wei, Dou Hu, Wei Zhou, Zhaojuan Yue, and Songlin Hu. 2021. Towards Propagation Uncertainty: Edge-enhanced Bayesian Graph Convolutional Networks for Rumor 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)*. Association for Computational Linguistics, Online, 3845–3854. <https://doi.org/10.18653/v1/2021.acl-long.297>
- [538] Tianwen Wei, Liang Zhao, Lichang Zhang, Bo Zhu, Lijie Wang, Haihua Yang, Biye Li, Cheng Cheng, Weiwei Lü, Rui Hu, Chenxia Li, Liu Yang, Xilin Luo, Xuejie Wu, Lunan Liu, Wenjun Cheng, Peng Cheng, Jianhao Zhang, Xiaoyu Zhang, Lei Lin, Xiaokun Wang, Yutuan Ma, Chuanhai Dong, Yanqi Sun, Yifu Chen, Yongyi Peng, Xiaojuan Liang, Shuicheng Yan, Han Fang, and Yahui Zhou. 2023. Skywork: A More Open Bilingual Foundation Model. *arXiv preprint arXiv: 2310.19341* (2023).
- [539] Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, Tianxiang Hu, Shangjie Li, Binyuan Hui, Bowen Yu, Dayiheng Liu, Baosong Yang, Fei Huang, and Jun Xie. 2023. PolyLM: An Open Source Polyglot Large Language Model. *arXiv preprint arXiv: 2307.06018* (2023).
- [540] Zimian Wei, Hengyue Pan, Linbo Qiao, Xin Niu, Peijie Dong, and Dongsheng Li. 2022. Cross-Modal Knowledge Distillation in Multi-Modal Fake News Detection. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4733–4737. <https://doi.org/10.1109/ICASSP43922.2022.9747280>
- [541] Zeming Wei, Yifei Wang, and Yisen Wang. 2023. Jailbreak and Guard Aligned Language Models with Only Few In-Context Demonstrations. *arXiv preprint arXiv: 2310.06387* (2023).
- [542] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from Language Models. *arXiv preprint arXiv: Arxiv-2112.04359* (2021).
- [543] Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Anna Rogers, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William Isaac. 2023. Sociotechnical Safety Evaluation of Generative AI Systems. *arXiv preprint arXiv: 2310.11986* (2023).
- [544] Orion Weller, Marc Marone, Nathaniel Weir, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2023. "According to ..." Prompting Language Models Improves Quoting from Pre-Training Data. *arXiv preprint arXiv: 2305.13252* (2023).
- [545] BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Froberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rhea Harlman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesh Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Laval-lée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwaa, Amanpreet Singh, Anastasia Cheveleva, Ana-Laure Ligozat, Arjun Subramonian, Aurélie Névél, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behrooz, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguié, Thanh Le, Tobi Oyeade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Perinián,

- Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrmann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sanger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sangaroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Th eo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv preprint arXiv: 2211.05100* (2022).
- [546] Guangyang Wu, Weijie Wu, Xiaohong Liu, Kele Xu, Tianjiao Wan, and Wenyi Wang. 2023. Cheap-fake Detection with LLM using Prompt Engineering. *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (2023). <https://doi.org/10.1109/ICMEW59549.2023.00025>
- [547] Hanqian Wu, Xinwei Li, Lu Li, and Qipeng Wang. 2022. Propaganda Techniques Detection in Low-Resource Memes with Multi-Modal Prompt Tuning. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*. 01–06. <https://doi.org/10.1109/ICME52920.2022.9859642>
- [548] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Hong Lin. 2023. AI-Generated Content (AIGC): A Survey. *arXiv preprint arXiv: Arxiv-2304.06632* (2023).
- [549] Jiaying Wu and Bryan Hooi. 2023. DECOR: Degree-Corrected Social Graph Refinement for Fake News Detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Long Beach, CA, USA) (*KDD '23*). Association for Computing Machinery, New York, NY, USA, 2582–2593. <https://doi.org/10.1145/3580305.3599298>
- [550] Jiaying Wu and Bryan Hooi. 2023. Fake News in Sheep’s Clothing: Robust Fake News Detection Against LLM-Empowered Style Attacks. *arXiv preprint arXiv: 2310.10830* (2023).
- [551] Jiaying Wu, Shen Li, Ailin Deng, Miao Xiong, and Bryan Hooi. 2023. Prompt-and-Align: Prompt-Based Social Alignment for Few-Shot Fake News Detection. *arXiv preprint arXiv: 2309.16424* (2023).
- [552] Junfei Wu, Qiang Liu, Weizhi Xu, and Shu Wu. 2022. Bias Mitigation for Evidence-Aware Fake News Detection by Causal Intervention. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) (*SIGIR '22*). Association for Computing Machinery, New York, NY, USA, 2308–2313. <https://doi.org/10.1145/3477495.3531850>
- [553] Junfei Wu, Weizhi Xu, Qiang Liu, Shu Wu, and Liang Wang. 2022. Adversarial Contrastive Learning for Evidence-aware Fake News Detection with Graph Neural Networks. *arXiv preprint arXiv: 2210.05498* (2022).
- [554] Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Derek F. Wong, and Lidia S. Chao. 2023. A Survey on LLM-generated Text Detection: Necessity, Methods, and Future Directions. *arXiv preprint arXiv: 2310.14724* (2023).
- [555] Kangxi Wu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2023. LLMdet: A Large Language Models Detection Tool. *arXiv preprint arXiv: 2305.15004* (2023).
- [556] Kun Wu, Xu Yuan, and Yue Ning. 2021. Incorporating relational knowledge in explainable fake news detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 403–415.
- [557] Liang Wu, Fred Morstatter, Kathleen M Carley, and Huan Liu. 2019. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD explorations newsletter* 21, 2 (2019), 80–90.
- [558] Lianwei Wu, Yuan Rao, Haolin Jin, Ambreen Nazir, and Ling Sun. 2019. Different Absorption from the Same Sharing: Sifted Multi-task Learning for Fake News Detection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 4644–4653. <https://doi.org/10.18653/v1/D19-1471>
- [559] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation Framework. *arXiv preprint arXiv: 2308.08155* (2023).
- [560] Weiye Wu and Shaoshan Liu. 2023. A Comprehensive Review and Systematic Analysis of Artificial Intelligence Regulation Policies. *arXiv preprint arXiv: 2307.12218* (2023).
- [561] Xueqing Wu, Kung-Hsiang Huang, Yi Fung, and Heng Ji. 2022. Cross-document Misinformation Detection based on Event Graph Reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 543–558. <https://doi.org/10.18653/v1/2022.naacl-main.40>
- [562] Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal Fusion with Co-Attention Networks for Fake News Detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, Online, 2560–2569. <https://doi.org/10.18653/v1/2021.findings-acl.226>
- [563] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Qin Liu, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huan, and Tao Gui. 2023. The Rise and Potential of Large Language Model Based Agents: A Survey. *arXiv preprint arXiv: 2309.07864* (2023).
- [564] Rui Xia, Kaizhou Xuan, and Jianfei Yu. 2020. A State-independent and Time-evolving Network for Early Rumor Detection in Social Media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 9042–9051. <https://doi.org/10.18653/v1/2020.emnlp-main.727>
- [565] Madelyne Xiao and Jonathan Mayer. 2023. The Challenges of Machine Learning for Trust and Safety: A Case Study on Misinformation Detection. *arXiv preprint arXiv: 2308.12215* (2023).
- [566] Jianhui Xie, Song Liu, Ruixin Liu, Yinghong Zhang, and Yuesheng Zhu. 2021. SERN: Stance Extraction and Reasoning Network for Fake News Detection. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2520–2524. <https://doi.org/10.1109/ICASSP39728.2021.9414787>
- [567] Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. *arXiv preprint arXiv: 2306.13063* (2023).
- [568] Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. 2023. Are Large Language Models Really Good Logical Reasoners? A Comprehensive Evaluation and Beyond. *arXiv preprint arXiv: 2306.09841* (2023).
- [569] Guohai Xu, Jiayi Liu, Ming Yan, Haotian Xu, Jinghui Si, Zhuoran Zhou, Peng Yi, Xing Gao, Jitao Sang, Rong Zhang, Ji Zhang, Chao Peng, Fei Huang, and Jingren Zhou. 2023. CValues: Measuring the Values of Chinese Large Language Models from Safety to Responsibility. *arXiv preprint arXiv: 2307.09705* (2023).

- [570] Han Xu, Jie Ren, Pengfei He, Shenglai Zeng, Yingqian Cui, Amy Liu, Hui Liu, and Jiliang Tang. 2023. On the Generalization of Training-based ChatGPT Detection Methods. *arXiv preprint arXiv: 2310.01307* (2023).
- [571] Weizhi Xu, Qiang Liu, Shu Wu, and Liang Wang. 2023. Counterfactual Debiasing for Fact Verification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 6777–6789. <https://doi.org/10.18653/v1/2023.acl-long.374>
- [572] 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* (Virtual Event, Lyon, France) (WWW '22). Association for Computing Machinery, New York, NY, USA, 2501–2510. <https://doi.org/10.1145/3485447.3512122>
- [573] Yiheng Xu, Hongjin Su, Chen Xing, Boyu Mi, Qian Liu, Weijia Shi, Binyuan Hui, Fan Zhou, Yitao Liu, Tianbao Xie, Zhoujun Cheng, Siheng Zhao, Lingpeng Kong, Bailin Wang, Caiming Xiong, and Tao Yu. 2023. Lemur: Harmonizing Natural Language and Code for Language Agents. *arXiv preprint arXiv: 2310.06830* (2023).
- [574] Jun Yan, Vikas Yadav, Shiyang Li, Lichang Chen, Zheng Tang, Hai Wang, Vijay Srinivasan, Xiang Ren, and Hongxia Jin. 2023. Virtual Prompt Injection for Instruction-Tuned Large Language Models. *arXiv preprint arXiv: 2307.16888* (2023).
- [575] Borui Yang, Wei Li, Liyao Xiang, and Bo Li. 2023. Towards Code Watermarking with Dual-Channel Transformations. *arXiv preprint arXiv: 2309.00860* (2023).
- [576] Fan Yang, Shiva K. Pentala, Sina Mohseni, Mengnan Du, Hao Yuan, Rhema Linder, Eric D. Ragan, Shuiwang Ji, and Xia (Ben) Hu. 2019. XFake: Explainable Fake News Detector with Visualizations. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia (Eds.). ACM, 3600–3604. <https://doi.org/10.1145/3308558.3314119>
- [577] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. *arXiv preprint arXiv: 2304.13712* (2023).
- [578] Kai-Cheng Yang and Filippo Menczer. 2023. Anatomy of an AI-powered malicious social botnet. *arXiv preprint arXiv: 2307.16336* (2023).
- [579] Ruichao Yang, Jing Ma, Hongzhan Lin, and Wei Gao. 2022. A Weakly Supervised Propagation Model for Rumor Verification and Stance Detection with Multiple Instance Learning. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) (SIGIR '22). Association for Computing Machinery, New York, NY, USA, 1761–1772. <https://doi.org/10.1145/3477495.3531930>
- [580] Ruichao Yang, Xiting Wang, Yiqiao Jin, Chaozhuo Li, Jianxun Lian, and Xing Xie. 2022. Reinforcement Subgraph Reasoning for Fake News Detection. In *KDD*.
- [581] Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, and Huan Liu. 2019. Unsupervised fake news detection on social media: A generative approach. In *Proceedings of 33rd AAAI Conference on Artificial Intelligence*.
- [582] Sin-han Yang, Chung-chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. Entity-Aware Dual Co-Attention Network for Fake News Detection. In *Findings of the Association for Computational Linguistics: EACL 2023*. Association for Computational Linguistics, Dubrovnik, Croatia, 106–113. <https://aclanthology.org/2023.findings-eacl.7>
- [583] 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* (2023).
- [584] Xianjun Yang, Wei Cheng, Linda Petzold, William Yang Wang, and Haifeng Chen. 2023. DNA-GPT: Divergent N-Gram Analysis for Training-Free Detection of GPT-Generated Text. *arXiv preprint arXiv: 2305.17359* (2023).
- [585] Xiaoyu Yang, Yuefei Lyu, Tian Tian, Yifei Liu, Yudong Liu, and Xi Zhang. 2020. Rumor Detection on Social Media with Graph Structured Adversarial Learning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, Christian Bessiere (Ed.). ijcai.org, 1417–1423. <https://doi.org/10.24963/ijcai.2020/197>
- [586] Xianjun Yang, Liangming Pan, Xuandong Zhao, Haifeng Chen, Linda Petzold, William Yang Wang, and Wei Cheng. 2023. A Survey on Detection of LLMs-Generated Content. *arXiv preprint arXiv: 2310.15654* (2023).
- [587] Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin. 2023. Shadow Alignment: The Ease of Subverting Safely-Aligned Language Models. *arXiv preprint arXiv: 2310.02949* (2023).
- [588] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The Dawn of LLMs: Preliminary Explorations with GPT-4V(ision). *arXiv preprint arXiv: 2309.17421* (2023).
- [589] Zhiwei Yang, Jing Ma, Hechang Chen, Hongzhan Lin, Ziyang Luo, and Yi Chang. 2022. A Coarse-to-fine Cascaded Evidence-Distillation Neural Network for Explainable Fake News Detection. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2608–2621. <https://aclanthology.org/2022.coling-1.230>
- [590] Dongyu Yao, Jianshu Zhang, Ian G. Harris, and Marcel Carlsson. 2023. FuzzLLM: A Novel and Universal Fuzzing Framework for Proactively Discovering Jailbreak Vulnerabilities in Large Language Models. *arXiv preprint arXiv: 2309.05274* (2023).
- [591] Jing Yao, Xiaoyuan Yi, Xiting Wang, Jindong Wang, and Xing Xie. 2023. From Instructions to Intrinsic Human Values - A Survey of Alignment Goals for Big Models. *arXiv preprint arXiv: 2308.12014* (2023).
- [592] 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* (2023).
- [593] Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing Large Language Models: Problems, Methods, and Opportunities. *arXiv preprint arXiv: 2305.13172* (2023).
- [594] Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023. Cognitive Mirage: A Review of Hallucinations in Large Language Models. *arXiv preprint arXiv: 2309.06794* (2023).
- [595] Wentao Ye, Mingfeng Ou, Tianyi Li, Yipeng chen, Xuetao Ma, Yifan Yanggong, Sai Wu, Jie Fu, Gang Chen, Haobo Wang, and Junbo Zhao. 2023. Assessing Hidden Risks of LLMs: An Empirical Study on Robustness, Consistency, and Credibility. *arXiv preprint arXiv: 2305.10235* (2023).
- [596] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A Survey on Multimodal Large Language Models. *arXiv preprint arXiv: 2306.13549* (2023).
- [597] Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do Large Language Models Know What They Don't Know?. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, Toronto, Canada, 8653–8665. <https://doi.org/10.18653/v1/2023.findings-acl.551>
- [598] Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. 2023. Low-Resource Languages Jailbreak GPT-4. *arXiv preprint arXiv: 2310.02446* (2023).



- [599] Kiyoon Yoo, W. Ahn, Jiho Jang, and N. Kwak. 2023. Robust Multi-bit Natural Language Watermarking through Invariant Features. *Annual Meeting of the Association for Computational Linguistics* (2023). <https://doi.org/10.18653/v1/2023.acl-long.117>
- [600] Shehel Yoosuf and Yin Yang. 2019. Fine-Grained Propaganda Detection with Fine-Tuned BERT. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*. Association for Computational Linguistics, Hong Kong, China, 87–91. <https://doi.org/10.18653/v1/D19-5011>
- [601] Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making Retrieval-Augmented Language Models Robust to Irrelevant Context. *arXiv preprint arXiv: 2310.01558* (2023).
- [602] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2023. Ferret: Refer and Ground Anything Anywhere at Any Granularity. *arXiv preprint arXiv: 2310.07704* (2023).
- [603] Hamza Yousuf, Sander van der Linden, Luke Bredius, GA Ted van Essen, Govert Sweep, Zohar Preminger, Eric van Gorp, Erik Scherder, Jagat Narula, and Leonard Hofstra. 2021. A media intervention applying debunking versus non-debunking content to combat vaccine misinformation in elderly in the Netherlands: A digital randomised trial. *EClinicalMedicine* 35 (2021), 100881.
- [604] Jiahao Yu, Xingwei Lin, and Xinyu Xing. 2023. GPTFUZZER : Red Teaming Large Language Models with Auto-Generated Jailbreak Prompts. *arXiv preprint arXiv: 2309.10253* (2023).
- [605] Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, Chunyang Li, Zheyuan Zhang, Yushi Bai, Yantao Liu, Amy Xin, Nianyi Lin, Kaifeng Yun, Linlu Gong, Jianhui Chen, Zhili Wu, Yunjia Qi, Weikai Li, Yong Guan, Kaisheng Zeng, Ji Qi, Hailong Jin, Jinxin Liu, Yu Gu, Yuan Yao, Ning Ding, Lei Hou, Zhiyuan Liu, Bin Xu, Jie Tang, and Juanzi Li. 2023. KoLA: Carefully Benchmarking World Knowledge of Large Language Models. *arXiv preprint arXiv: 2306.09296* (2023).
- [606] Peipeng Yu, Jiahua Chen, Xuan Feng, and Zhihua Xia. 2023. CHEAT: A Large-scale Dataset for Detecting ChatGPT-writteN AbsTracts. *arXiv preprint arXiv: Arxiv-2304.12008* (2023).
- [607] Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2020. A Survey of Knowledge-Enhanced Text Generation. *arXiv preprint arXiv: 2010.04389* (2020).
- [608] Zihan Yu, Liang He, Zhen Wu, Xinyu Dai, and Jiajun Chen. 2023. Towards Better Chain-of-Thought Prompting Strategies: A Survey. *arXiv preprint arXiv: 2310.04959* (2023).
- [609] Chunyuan Yuan, Qianwen Ma, Wei Zhou, Jizhong Han, and Songlin Hu. 2019. Jointly Embedding the Local and Global Relations of Heterogeneous Graph for Rumor Detection. In *2019 IEEE International Conference on Data Mining (ICDM)*. 796–805. <https://doi.org/10.1109/ICDM.2019.00090>
- [610] Chunyuan Yuan, Qianwen Ma, Wei Zhou, Jizhong Han, and Songlin Hu. 2020. Early Detection of Fake News by Utilizing the Credibility of News, Publishers, and Users based on Weakly Supervised Learning. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 5444–5454. <https://doi.org/10.18653/v1/2020.coling-main.475>
- [611] Lifan Yuan, Yangyi Chen, Xingyao Wang, Yi R. Fung, Hao Peng, and Heng Ji. 2023. CRAFT: Customizing LLMs by Creating and Retrieving from Specialized Toolsets. *arXiv preprint arXiv: 2309.17428* (2023).
- [612] Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Pinjia He, Shuming Shi, and Zhaopeng Tu. 2023. GPT-4 Is Too Smart To Be Safe: Stealthy Chat with LLMs via Cipher. *arXiv preprint arXiv: 2308.06463* (2023).
- [613] Zhenrui Yue, Huimin Zeng, Ziyi Kou, Lanyu Shang, and Dong Wang. 2022. Contrastive Domain Adaptation for Early Misinformation Detection: A Case Study on COVID-19. In *Proc. of CIKM*.
- [614] Zhenrui Yue, Huimin Zeng, Yang Zhang, Lanyu Shang, and Dong Wang. 2023. MetaAdapt: Domain Adaptive Few-Shot Misinformation Detection via Meta Learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 5223–5239. <https://doi.org/10.18653/v1/2023.acl-long.286>
- [615] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending Against Neural Fake News. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 9051–9062. <https://proceedings.neurips.cc/paper/2019/hash/3e9f0fc9b2f89e043bc6233994dfcf76-Abstract.html>
- [616] Fengzhu Zeng and Wei Gao. 2022. Early Rumor Detection Using Neural Hawkes Process with a New Benchmark Dataset. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 4105–4117. <https://doi.org/10.18653/v1/2022.naacl-main.302>
- [617] Fengzhu Zeng and Wei Gao. 2023. Prompt to be Consistent is Better than Self-Consistent? Few-Shot and Zero-Shot Fact Verification with Pre-trained Language Models. *arXiv preprint arXiv: 2306.02569* (2023).
- [618] Bohui Zhang, Ioannis Reklou, Nitisha Jain, Albert Meroño Peñuela, and Elena Simperl. 2023. Using Large Language Models for Knowledge Engineering (LLMKE): A Case Study on Wikidata. *arXiv preprint arXiv: 2309.08491* (2023).
- [619] Chaoning Zhang, Chenshuang Zhang, Chenghao Li, Yu Qiao, Sheng Zheng, Sumit Kumar Dam, Mengchun Zhang, Jung Uk Kim, Seong Tae Kim, Jinwoo Choi, Gyeong-Moon Park, Sung-Ho Bae, Lik-Hang Lee, Pan Hui, In So Kweon, and Choong Seon Hong. 2023. One Small Step for Generative AI, One Giant Leap for AGI: A Complete Survey on ChatGPT in AIGC Era. *arXiv preprint arXiv: 2304.06488* (2023).
- [620] Huaiwen Zhang, Quan Fang, Shengsheng Qian, and Changsheng Xu. 2019. Multi-modal Knowledge-aware Event Memory Network for Social Media Rumor Detection. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21–25, 2019*. 1942–1951. <https://doi.org/10.1145/3343031.3350850>
- [621] Hangfan Zhang, Zhimeng Guo, Huaisheng Zhu, Bochuan Cao, Lu Lin, Jinyuan Jia, Jinghui Chen, and Dinghao Wu. 2023. On the Safety of Open-Sourced Large Language Models: Does Alignment Really Prevent Them From Being Misused? *arXiv preprint arXiv: 2310.01581* (2023).
- [622] Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023. Extractive Summarization via ChatGPT for Faithful Summary Generation. *arXiv preprint arXiv: 2304.04193* (2023).
- [623] Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023. How Language Model Hallucinations Can Snowball. *arXiv preprint arXiv: 2305.13534* (2023).
- [624] Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. 2023. InternLM-XComposer: A Vision-Language Large Model for Advanced Text-image Comprehension and Composition. *ArXiv preprint abs/2309.15112* (2023). <https://arxiv.org/abs/2309.15112>
- [625] Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. 2023. Retrieve Anything To Augment Large Language Models. *arXiv preprint arXiv: 2310.07554* (2023).
- [626] Shuo Zhang, Liangming Pan, Junzhou Zhao, and William Yang Wang. 2023. Mitigating Language Model Hallucination with Interactive Question-Knowledge Alignment. *arXiv preprint arXiv: 2305.13669* (2023).

- [627] Tianhua Zhang, Hongyin Luo, Yung-Sung Chuang, Wei Fang, Luc Gaitskell, Thomas Hartvigsen, Xixin Wu, Danny Fox, Helen Meng, and James Glass. 2023. Interpretable Unified Language Checking. *arXiv preprint arXiv: Arxiv-2304.03728* (2023).
- [628] Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3Exam: A Multilingual, Multimodal, Multilevel Benchmark for Examining Large Language Models. *arXiv preprint arXiv: 2306.05179* (2023).
- [629] Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining Dual Emotion for Fake News Detection. In *Proc. of WWW*.
- [630] Xuan Zhang and Wei Gao. 2023. Towards LLM-based Fact Verification on News Claims with a Hierarchical Step-by-Step Prompting Method. *arXiv preprint arXiv: 2310.00305* (2023).
- [631] Xichen Zhang and Ali A Ghorbani. 2020. An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management* 57, 2 (2020), 102025.
- [632] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren’s Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *arXiv preprint arXiv: 2309.01219* (2023).
- [633] Yinuo Zhang, Zhulin Tao, Xi Wang, and Tongyue Wang. 2023. INO at Factify 2: Structure Coherence based Multi-Modal Fact Verification. *arXiv preprint arXiv: 2303.01510* (2023).
- [634] Yizhou Zhang, Loc Trinh, Defu Cao, Zijun Cui, and Yan Liu. 2023. Detecting Out-of-Context Multimodal Misinformation with interpretable neural-symbolic model. *arXiv preprint arXiv: Arxiv-2304.07633* (2023).
- [635] Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. 2023. How Do Large Language Models Capture the Ever-changing World Knowledge? A Review of Recent Advances. *arXiv preprint arXiv: 2310.07343* (2023).
- [636] Zhixin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2023. SafetyBench: Evaluating the Safety of Large Language Models with Multiple Choice Questions. *arXiv preprint arXiv: 2309.07045* (2023).
- [637] Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, and Shafiq Joty. 2023. Retrieving Multimodal Information for Augmented Generation: A Survey. *arXiv preprint arXiv: 2303.10868* (2023).
- [638] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models. *arXiv preprint arXiv: Arxiv-2303.18223* (2023).
- [639] Xuandong Zhao, Prabhajan Ananth, Lei Li, and Yu-Xiang Wang. 2023. Provable Robust Watermarking for AI-Generated Text. *arXiv preprint arXiv: 2306.17439* (2023).
- [640] Jiaqi Zheng, Xi Zhang, Sanchuan Guo, Quan Wang, Wenyu Zang, and Yongdong Zhang. 2022. MFAN: Multi-modal Feature-enhanced Attention Networks for Rumor Detection. *IJCAI*.
- [641] Kaizhi Zheng, Xuehai He, and Xin Eric Wang. 2023. MiniGPT-5: Interleaved Vision-and-Language Generation via Generative Vokens. *arXiv preprint arXiv: 2310.02239* (2023).
- [642] Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, Limao Xiong, Lu Chen, Zhiheng Xi, Nuo Xu, Wenbin Lai, Minghao Zhu, Cheng Chang, Zhangyue Yin, Rongxiang Weng, Wensen Cheng, Haoran Huang, Tianxiang Sun, Hang Yan, Tao Gui, Qi Zhang, Xipeng Qiu, and Xuanjing Huang. 2023. Secrets of RLHF in Large Language Models Part I: PPO. *arXiv preprint arXiv: 2307.04964* (2023).
- [643] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinu Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less Is More for Alignment. *arXiv preprint arXiv: 2305.11206* (2023).
- [644] 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*. 1–20.
- [645] Wangchunshu Zhou, Yuchen Eleanor Jiang, Long Li, Jialong Wu, Tiannan Wang, Shi Qiu, Jintian Zhang, Jing Chen, Ruipu Wu, Shuai Wang, Shiding Zhu, Jiyu Chen, Wentao Zhang, Ningyu Zhang, Hua-jun Chen, Peng Cui, and Mrinmaya Sachan. 2023. Agents: An Open-source Framework for Autonomous Language Agents. *arXiv preprint arXiv: 2309.07870* (2023).
- [646] Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)* 53, 5 (2020), 1–40.
- [647] Chenguang Zhu, Yichong Xu, Xiang Ren, Bill Yuchen Lin, Meng Jiang, and Wenhao Yu. 2022. Knowledge-Augmented Methods for Natural Language Processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*. Association for Computational Linguistics, Dublin, Ireland, 12–20. <https://doi.org/10.18653/v1/2022.acl-tutorials.3>
- [648] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv preprint arXiv: Arxiv-2304.10592* (2023).
- [649] Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, and Xing Xie. 2023. PromptBench: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts. *arXiv preprint arXiv: 2306.04528* (2023).
- [650] Yongchun Zhu, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, and Fuzhen Zhuang. 2022. Generalizing to the Future: Mitigating Entity Bias in Fake News Detection. In *Proc. of SIGIR*.
- [651] Yongchun Zhu, Qiang Sheng, Juan Cao, Qiong Nan, Kai Shu, Minghui Wu, Jindong Wang, and Fuzhen Zhuang. 2022. Memory-Guided Multi-View Multi-Domain Fake News Detection. *IEEE Transactions on Knowledge and Data Engineering* (2022). <https://doi.org/10.1109/TKDE.2022.3185151>
- [652] Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Exploring AI Ethics of ChatGPT: A Diagnostic Analysis. *arXiv preprint arXiv: Arxiv-2301.12867* (2023).
- [653] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. 2023. Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv preprint arXiv: 2310.01405* (2023).
- [654] Anni Zou, Zhuosheng Zhang, and Hai Zhao. 2023. Decker: Double Check with Heterogeneous Knowledge for Commonsense Fact Verification. *Annual Meeting of the Association for Computational Linguistics* (2023). <https://doi.org/10.48550/arXiv.2305.05921>
- [655] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and Resolution of Rumours in Social Media: A Survey. *ACM Comput. Surv.* 51, 2 (2018), 32:1–32:36. <https://doi.org/10.1145/3161603>
- [656] Yuhui Zuo, Wei Zhu, and Guoyong GUET Cai. 2022. Continually Detection, Rapidly React: Unseen Rumors Detection Based on Continual Prompt-Tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 3029–3041.