# **Final Project Related Work**

Title:

Leveraging LLMs for LLM-Generated Fake News Detection: Insights from COVID-19 Misinformation

Citation MLA 9:

Hong, Dao Ngoc, et al. “Leveraging Llms for LLM-generated fake news detection: Insights from covid-19 misinformation.” *2024 IEEE 16th International Conference on Computational Intelligence and Communication Networks (CICN)*, 22 Dec. 2024, pp. 1460–1466, <https://doi.org/10.1109/cicn63059.2024.10847475>.

Summary:

The paper sets out to answer two questions: Can LLMs detect fake news from a pool of fake and valid news that the LLM itself created? How consistent are LLMs in evaluating tweets generated by other LLMs? Using GPT-3.5, GPT-4, Gemini, and Claude, each model was tasked with reading the "COVID-19 Fake News Dataset" and generating tweets based on what was read. After, the models were tasked with evaluating the generated tweets and classifying each as valid or fake news. After the experiment was conducted, it turned out that each model used was discovered to be better than others at certain metrics and uses. For instance, GPT-3.5 had the best precision scores, meaning it was less prone to label fake news as true, while Gemini had a higher recall score, meaning it was useful for tasks related to identifying larger portions of fake news instances.

Things I’ve learned:

From this paper, I learned a few things that would help me with my approach in my project. One is that I need to double check my dataset(s) for the use of "Partially False" data. I, like the authors of the paper, would like to reduce the ambiguity and complexity of the analysis that needs to be conducted. The other thing I learned is that semantic simplicity and/or clarity will have an impact on my model's performance and may need to be mentioned in the findings. I'm curious to know, with the datasets I use, if more reputable media outlets will drive more accurate scores compared to less reputable sources that may have a lower standard of writing quality.

Title:

Assessing the Effectiveness of GPT-3 in Detecting False Political Statements: A Case Study on the LIAR Dataset

Citation MLA 9:

Buchholz, Mars. “Assessing the Effectiveness of GPT-3 in Detecting False Political Statements: A Case Study on the LIAR Dataset.” *Arxiv*, 2023.

Summary:

This paper looks at using GPT-3 to spot false political statements using the LIAR dataset (which has statements rated on a six point truthfulness scale by PolitiFact). Researchers tried two different approaches: fine tuning GPT-3 models and using zero shot learning with prompt engineering. Their fine tuned Curie model hit 29.5% accuracy which beat previous leading models that relied on metadata and linguistic patterns, while only using the statement text itself.

Things I've learned:

This research shows that modern LLMs can perform well at fake news detection without needing extra metadata or complex linguistic feature extraction. This could be useful for my project as I am trying to keep the complexity of it to a minimum ion order to get the model to work. I also like how their zero shot approach showed LLMs can explain their reasoning along with the classifications that could make my fake news detector more transparent for users. I would like to explore this feature opportunity further.

Title:

Disinformation Capabilities of Large Language Models

Citation MLA 9:

Vykopal, Ivan, et al. “Disinformation capabilities of large language models.” *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 14830–14847, <https://doi.org/10.18653/v1/2024.acl-long.793>.

Summary:

This study by Vykopal and others looks at 10 different LLMs handle generating fake news when prompted to create false articles wtih 20 different disinformation topics. They looked at how convincing the fake news was, whether the LLMs agreed or disagreed with false narratives, how often they gave safety warnings, and how easy it was to spot the AI-generated content. There were huge differences between models. Some like Vicuna and GPT-3 Davinci had no problem creating fake news, while others like Falcon usually refused or added disclaimers stating "this is false."

Things I've learned:

A takeaway for me from this research for my project is how differently various LLMs handle safety measures. It's encouraging that existing detection methods can spot machine generated fake news with decent accuracy (0.8 F1 score for the best detectors). In a perfect world O might be able to use similar approaches. I really liked their method of evaluating both the content agreement and writing style, which gives me ideas for my own evaluation metrics. Their "safe/dangerous" categorization system was also very interesting to learn about even though I will not be applying it to my project.

Title:

CAN LLM-GENERATED MISINFORMATION BE DETECTED?

Citation MLA 9:

Chen, Canyu, and Kai Shu. “CAN LLM-GENERATED MISINFORMATION BE DETECTED?” *Arxiv*, 2024.

Summary:

This paper by Chen and Shu looks at whether fake news created by LLMs is harder to catch than human written fake news. They break down LLM misinformation into three types: unintentional hallucinations (when LLMs just make stuff up), arbitrary generation (when they're told to create fake news), and controllable generation (where they rewrite existing misinformation while keeping the same meaning). After a good number of tests, they found that LLM generated fake news that keeps the same meaning as human-written fake news is more difficult for both humans and automated systems to detect. Basically, LLMs can create more believable lies, which is pretty concerning.

Things I've learned:

This research is super relevant for my fake news detector since it shows LLM generated misinformation is a special kind of challenge because of how convincing it can be. It was in fact that even the best detectors struggle with LLMs generated content suggests I may be in for a rude awakening when trying to finetuning my model to spot misinformation.

Title:

Towards Reliable Misinformation Mitigation: Generalization, Uncertainty, and GPT-4

Citation MLA 9:

Pelrine, Kellin, et al. “Towards reliable misinformation mitigation: Generalization, uncertainty, and GPT-4.” *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, <https://doi.org/10.18653/v1/2023.emnlp-main.395>.

Summary:

This paper explores how GPT-4 compares against older methods for spotting misinformation. The researchers found GPT-4 beats previous approaches across different settings and languages, but they went beyond just basic classification accuracy. They focused on two key problems: generalization (how well methods work on new types of content) and uncertainty handling. They came up with techniques to identify "impossible" examples (statements that don't have enough context to verify) and showed huge performance improvements when models are allowed to say "I don't know" instead of forcing a true/false judgment. They also released a new dataset called LIAR-New with paired English am dFrench fake news data and labels indicating whether statements have enough context to be properly evaluated.

Things I've learned:

My biggest take away from this research is how important it is to let fake news detectors express uncertainty. Instead of forcing every piece of content into either "true" or "false" buckets, letting the model say "I don't have enough context to tell" can make the whole system way more reliable. This makes a ton of sense in the real world since saying "I don't know" is usually better than confidently giving the wrong answer (false positives). For my project, I'm going to try and implement some uncertainty aware feature that can flag statements lacking enough context for verification. This should make my system both perform better and be more trustworthy compared to models that force a judgment on everything and decrease from their overall performance (think F1- scores, precision, etc.).

Title:

Generative Large Language Models in Automated Fact-Checking: A Survey

Citation MLA 9:

Vykopal, Ivan, et al. "Generative Large Language Models in Automated Fact-Checking: A Survey." arXiv, 30 Oct. 2024, arXiv:2407.02351v2.

Summary:

We aim to present an overview of techniques employed in fact-checking using generative LLMs.

Things I've learned:

Fact-Checking Tasks, ***specifically how many paper’s used these methods for fact checking***

* 3.1 Fact Verification & Fake News Detection
* 3.2 Evidence Retrieval
* 3.3 Claim Detection
* 3.4 Previously Fact-Checked Claims Detection

Taxonomy of Methods, ***specifically discuss the methods used to use to perform fact checking tasks and categorize methods based on their output type..***

* 4.1 Structured Output
  + further categorize these methods into three output types: classification, regression and ranking.
    - Classification
    - Regression
    - Ranking
* 4.2 Unstructured Output
  + Explanation Generation
  + Claim and Question Generation
  + Query GeneratioN
* 4.3 Synthetic Data Generation

Techniques, ***specifically analyze how different studies applied techniques to guide LLMs in achieving accurate outputs.***

* 5.1 Prompting
* 5.2 Fine-Tuning
* 5.3 Augmentation with External Knowledge

LLM Pipelines, specifically

Languages