# Fact-checking Tweets using Retrieval Augmented Large Language Model

Millennium Bismay
*UIN: 633002191*
*mbismay@tamu.edu*
*Department of Computer Science and Engineering*
*Texas A&M University*
*College Station, TX, 77840*

Anushka Garg
*UIN: 532008694*
*anushkagarg@tamu.edu*
*Department of Computer Science and Engineering*
*Texas A&M University*
*College Station, TX, 77840*

***Abstract*—Automatic detection of fake content in social media, especially Twitter(now X) is a persistent problem. In theory, identifying false information on social networking sites is a binary classification issue. But the sheer amount of daily tweets would make it impossible to manually fact-check even a tiny portion of them. To address this challenge, the team behind the Truthseeker dataset crawled and crowd-sourced one of the most extensive ground-truth datasets containing more than 180,000 labels from 2009 to 2022 for tweets with a 5-label and 3-label classification using Amazon Mechanical Turk. However, it is impossible to perform this activity in near real-time. We propose a Large Language Model based approach which fact-checks the truthfulness of the tweet by comparing it with legitimate news sources of corresponding topics in real time via Retrieval Augmented Generation. We aim to build a system that is faithful to the legitimate news-source to generate a truthfulness value for every tweet.**

**Index Terms - Fake News Detection, Automatic Detection, Retrieval Augmented Generation, Large Language Model, Faithful AI**

**** You can find the code and video at [Code][Video]**

## 1. Introduction

Social media usage has become a need for human survival in the modern day. There are several benefits for both individuals and businesses as a consequence of social media's exponential development in popularity and usage. In addition to offering leisure and pleasure, social media platforms enable users to share their own material and reach a large audience for the consumption of a variety of information, including news from around the world and locally. Social media's widespread use has changed the nature of communication by establishing a common platform that supports a wide variety of user interactions and behaviors. Social media has made it simpler to spread false news, which makes it possible for inaccurate or misleading information to spread fast to a big audience. During the 2016 US presidential election, research showed that approximately 14% of Americans relied on social media as their primary news source, surpassing print and radio. One major challenge for analyzing social media platforms is collecting and labeling a large enough training dataset to be used as ground truth [1].

Social networking sites offer many benefits, but they are also important sources of erroneous or misleading information. Every day, a great deal of false information is shared on social media, which might have negative effects on people's lives and society as a whole. Misinformation propagated via social media has far-reaching effects that may greatly affect public opinion, political consequences, and decision-making. For example, studies show that during the 2016 U.S. presidential election, social media surpassed traditional print and radio outlets to become a significant news source for around 14% of Americans. Thus, it's critical to investigate practical strategies for locating and preventing the propagation of false information on social media platforms.

A research [2] found that false news about the two presidential candidates, Donald Trump and Hillary Clinton, was shared millions of times on social media. Likewise, in the 2021 US presidential election campaign, recent research discovered more extensive misinformation campaigns around COVID-19. Moreover, in the aftermath of the 2021 election, specific security associations caught fake news campaigns claiming election fraud detected. These examples show that methods for identifying fake news are a relevant research topic and a pressing societal need. While different issues regarding tweet classification, such as topic or sentiment detection, are considerably researched, automatic fake news detection requires more engagement [3].

Most of the work in this domain has been performed on statistical analysis of tweets and users who posted the tweets. The most important factor in determining an ML/DL model's reliability and validity is its dataset. Nonetheless, it is certain that the current databases

of false news have limits. To reflect the sophisticated creation patterns of the new fake news producers, the majority of the datasets that are currently in use need to be updated. Furthermore, several social media users and messages are inaccessible after being identified as potentially harmful or dubious. The applicability of any model on new data input cannot be guaranteed by high performance on such a dataset. One major work has been done by the group behind the Truthseeker dataset [4] which emphasises on the truthfulness of the tweets as compared to the news from Politifact. They crawled and crowd-sourced one of the most extensive ground-truth datasets containing more than 180,000 labels from 2009 to 2022 for tweets with a 5-label and 3-label classification using Amazon Mechanical Turk. They employed highly skilled Amazon Mechanical Turkers whose main tasks are two fold - (1) identifying and understanding the truthful news related to the tweet and (2) deciding the truthfulness of the tweet as compared to the correct news. However, the manual effort is tedious and not real-time or near-real-time making the fake news to propagate and be flagged at a later point of time

Fake news has a temporal bias and has been found to be travelling at a faster speed and is capable to reaching more number of people in a small amount of time and smaller number of hops. Hence, it is crucial to identify a fake tweet at the earliest and limit its propagation. Our method proposes a framework to flag a fake tweet in near real-time so that it can be purged or stopped from propagating -

- Identify the topics a tweet is about and crawl the internet for the most recent news about the topics from legitimate sources. We have defined certain sources as being legitimate, such as unbiased news bodies like CNN, Politifact, BBC, etc.
- Retrieve the most recent news from multiple legitimate sources, summarize the news and generate a faithful summary using Large Language Model. This mimics the behaviour of the Amazon Mechanical Turkers employed by the Truthseeker Team who understand and faithfully accumulates the world knowledge.
- Compare the faithful summary with the tweet to predict the truthfulness of the tweet. This mimics the decisiveness behaviour of the Amazon Mechanical Turkers employed by the Truthseeker Team.

## 2. Related Works

It is imperative to accurately detect fake news, and a trustworthy dataset is a key tool in this process. However, it becomes difficult to train algorithms that can correctly detect false news in the absence of a pertinent and comprehensive dataset. The authors of [5] talk about how people are becoming more interested in identifying and confirming the legitimacy of material pertaining to false news. They carried out an extensive analysis of 118 online datasets that are openly accessible. The datasets were divided into groups according to how much of an emphasis they placed on identifying satire, fact-checking, evaluating, and spotting false news. Additionally, the researchers looked at each dataset's features and applications, pointing out problems and areas that needed more investigation. The construction of truth-based datasets has been an endeavor undertaken for many years. One of the earliest examples of combining truth scores from multiple sources is the original Politifact dataset [6]. This dataset merged the truth scores from two websites, Channel 4's fact-checking blog and the Truth- O-Meter from Politifact, into a single scale that included five labels: True, Mostly True, HalfTrue, Mostly False, and False. The dataset also includes the URLs and scores of the news.

The Truthseeker dataset [4] creation process relied on the Politifact's 5-label structure and a combination of expert and crowdsourced data crawling to balance qualitative and quantitative data, which is crucial for creating datasets for models to train on efficiently. From the statements collected by crowd-sourcing, 2-5 keywords were extracted manually with the label to be True or False, denoting whether a statement was True or False. The tweeter was crawled from 2009-2022, to get unique tweets from these keywords and a dataset was prepared. This was provided to the Amazon Mechanical Turkers to validate. They were to choose between 5 categories - Agree, Mostly Agree, Unknown, Mostly Disagree, Disagree for a tweet whether it was true to the statement. These Amazon Mechanical Turkers were highly skilled and each tweet was sent to 5 of them in order to create a majority voting. There were two types of Majority voting produced in the dataset -

- **5-Way Majority**: The final label was decided by the label which has the highest vote from the above mentioned categories. If 'Unknown' was the highest category then the label was 'NO MAJORITY'.
- **3-Way Majority**: The final label was decided by clubbing the Agree and Mostly Agree categories to Agree, and Disagree and Mostly Disagree categories to Disagree. Again, if 'Unknown' was the highest category then the label was 'NO MAJORITY'.

They also provided statistical features from the user who posted the tweet as well as tweet features. The data cleaning, feature extraction, and modelling for our work will be discussed in the next section. A complete statistical approach has been used in [7] which scrutinizes the effectiveness and dynamics of fact-checking in countering misinformation associated with the hashtag. The authors delve into the prevalence and nature of false information, examining how users engage with and disseminate fact-checked content.

The work [8] explores the potential of Large Language Models, like GPT-3, in generating synthetic data that faithfully replicates real-world social science datasets. The study aims to assess the accuracy of synthetic data in preserving statistical and structural characteristics of the original data, enabling researchers to use it for analysis without compromising privacy or confidentiality. The research investigates the utility of synthetic data generated by language models for computational social science research, offering insights into the practical applications of such synthetic datasets. This inspired our work that Large Language Models can faithfully summarize the news from legitimate sources. On a similar line, another research on Measuring Faithfulness using Chain-of-Thought [9] proposes a novel methodology to assess how well a model maintains consistency and coherence in its responses across a sequence of queries. By introducing metrics to quantify faithfulness, the study aims to enhance the evaluation of natural language generation models, particularly in scenarios requiring logical continuity. This research contributes to advancing the understanding and measurement of faithfulness in language models, offering valuable insights into the capabilities and limitations of these models in sustaining coherent reasoning over multiple turns of conversation.

It has been also found that Large Language Models have an emergent ability [10] which allows them to perform In-context tasks by zero-shot or few-shot settings. We aim to use this for topic modelling to extract important keywords from the tweet. Kai et al. showed that the performance of Large Language Models improved significantly as opposed to their zero-shot ability when fine tuned with data from a certain domain [11]. LoRA [12] has been one of the cruicial method for finetuning where the authors have shown that Low Rank adaptation of the Update matrix is much more beneficial than full finetuning for multiple downstream tasks. In the paper, Chain-of-Thought Prompting Elicits Reasoning in Large Language Models [13], it was illustrated how reasoning capability of LLMs can be enhanced highly by using Chain-of-Thought, which we will be using for predicting the truthfulness value of the tweets.

## 3. Methodology

The Truthseeker dataset [4] has been instrumental to our research. We researched on multiple datasets, most of which focused on certain statistical aspects of tweets or the sentiment and content of tweets. But none of the work was exhaustive about the truthfulness of the tweet with relevance to the faithful news source. It is the first time we have a human annotated large scale truthfulness indicator for tweet against faithful news articles. As of this checkpoint, we have worked on data cleaning, preprocessing, and developing baseline models.

### 3.1. Data Processing

As explained in Section 2, the most important features of the Truthseeker datasets are the highly skilled Amazon Mechanical Turkers' annotated labels for every tweet as compared to their relevance to the statements from the Politifact dataset [6]. However, there were around 20% of the data without any absolute majority for the 5-way label, i.e. 'NO MAJORITY', hence we removed such data from our work.

The tweets contain a lot of jargons, usernames, emails, and urls, which were cleaned as part of data cleaning. We kept the tense intact as we observed that Large Language Models tend to give better results for keywords of the tweets with a more intelligible structure, mostly because of the way LLMs are trained. Topic Modelling however is a work in progress and will be completed in next iteration.

The dataset has 2 labels, 5-way label and 3-way labels. Hence, we split it into two datasets, each with one type of labels. This is to understand the behaviour of models with understanding high level and low level data. We will call the 3-way labels as high level data as it is not that granular, i.e. Agree and Disagree. We will call the 5-way labels as low level data as it is granular, i.e. Agree, Mostly Agree, Mostly Disagree, Disagree. Once we have these two datasets, combined with the target values of the statements, i.e. True and False, we derive the Truthfulness labels which will be our final label used for predictions. We will use it as predictions for statistical modelling and well as for finetuning the LLM using the LoRA method. For the high label dataset (with 3 - way label), we followed the following heuristics to get the final Truthfulness Labels - Similarly we followed the following heuristics to get the final Truthfulness Labels -

TABLE 1. Truthfulness Classification (3 - way)

| **Statement (T/F)** | **3 - way Majority Answer** | **Truthfulness** |
|---|---|---|
| T | Agree | True |
| T | Disagree | False |
| F | Agree | False |
| F | Disagree | True |

TABLE 2. Truthfulness Classification (5 - way)

| **Statement (T/F)** | **5 - way Majority Answer** | **Truthfulness** |
|---|---|---|
| T | Agree | True |
| T | Mostly Agree | Mostly True |
| T | Mostly Disagree | Mostly False |
| T | Disagree | False |
| F | Agree | False |
| F | Mostly Agree | Mostly False |
| F | Mostly Disagree | Mostly True |
| F | Disagree | True |

So, finally we have 2 datasets with Truthfulness labels with different levels of granularity.

### 3.2. Statistical Modelling

The Truthseeker dataset has provided with multiple statistical features to perform statistical modelling. We can use them to perform an initial level of prediction whether a tweet is truthful or not. In application, we can use it to flag a tweet temporarily before a final Truthfulness value is predicted by RAG using LLM. We can broadly categorize these features into three categories -

- **Text Features:** Average word length, Percent of text including spaCy tags for ORG, PERSON, MONEY, DATA, PRODUCT, EVENT, etc.
- **Lexical Features:** Number of verbs, adjectives, pronouns, commas, dots, punctuations, capitalized letters, digits, etc.
- **Meta-Deta Features:** Number of followers, friends, tweets, mentiones, replies, quotes, etc.

We performed extensive hyperparameter tuning for the models - Random Forest, AdaBoost, XGBoost, and LightGBM. The model evaluation results are showcased in Table - 3. We used the model performance by the Truthseeker team as our baseline.

TABLE 3. Model Evaluation Metrics

| Model | Developed by | Precision | Recall | F-1 | Accuracy |
|---|---|---|---|---|---|
| AdaBoost | Truthseeker | 0.60 | 0.60 | 0.60 | 0.60 |
| AdaBoost | Ours | 0.66 | 0.65 | 0.65 | 0.65 |
| Random Forest | Both | 0.70 | **0.70** | **0.70** | **0.70** |
| XGBoost | Ours | 0.70 | 0.68 | 0.69 | **0.70** |
| **LightGBM** | Ours | **0.71** | 0.69 | **0.70** | **0.70** |

Given our usecase, it is critical to flag a false tweet as soon as possible, which means we need as low False Positives as possible, which implies we need a high Precision. Our AdaBoost Model performed better than the Truthseeker team's AdaBoost almost in all aspects. We match the performance of Random Forest. We outperformed all statistical models by our LightGBM model with the highest precision of 0.71 which is a slight improvement, but we observed overfitting of statistical model beyond that point. Given the high importance of Precision, we clearly find our LightGBM model to be the best of all, beating the Truthseeker team's models, with a high Precision and equal F-1 and Accuracy scores.

### 3.3. Retrieval augmented LLM (RaLLM)

The feature importance of statistical model shows that most of the top - 10 important features are not related to the tweets, instead are related to the user profile such as *friends_count, statuses_count, cred, influence*, etc. These features could be helpful in identifying a fake profile or anomalous profile but identifying whether a tweet is truthful or not, needs attention to the real - time world news. It has recently been found that, number of fake tweets have shooted up in the last couple of years and the proportion of fake tweets from verified users have also sky rocketed. This tells that the statistical methods might not be really helpful in the near future. Hence, we need to look at an alternative, which should identify the fake tweets in real-time autonomously. We follow the framework in Figure 1

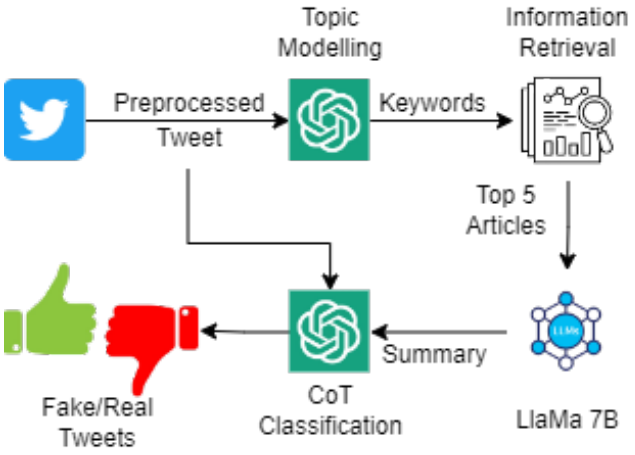


Figure 1. Retrieval Augmented LLM Framework

**3.3.1. Topic Modeling.** We used the zero shot ability of ChatGPT to extract 3 - 5 important keywords from the cleaned tweets as showin in Figure 1. We used a system prompt to ask ChatGPT to provide with keywords. We provide a system prompt which asks ChatGPT to be unbiased and provide a list of keywords in a definite format as shown below.

**System Prompt:** *You are an expert journalist. You do not hallucinate. Your views are not aligned to any political party or propaganda. You are unbiased, neutral and faithful to the text provided. List 3 to 5 most important topics or keywords discussed in the following sentences in the format [keyword1, keyword2, keyword3]-*

**Content:** *Processed Tweet*

**Response:** *keyword1, keyword2, keyword3, keyword4*

**3.3.2. Top - 5 Article Retrieval and Summarization.** We used *BeautifulSoup* to scrap the top-5 news articles from internet from the recognised sources like BBC, CNN, and Politifact as shown in Figure 1. We processed the text to remove any strong, critical or biased words. Then we use ChatGPT again to summarize the collective news in as unbiased, neutral and faithful way as possible.

**3.3.3. Chain-of-Thought Faithful Reasoning Generation.** Here we use the Prompt Engineering to its fullest. We have the processed tweet and the summary from the top-5 articles, We ask ChatGPT to generate the chain of thought behind whether the tweet is truthful or not, to the summary of articles as shown in Figure 2. In the paper, Chain-of-Thought Prompting Elicits Reasoning in Large Language Models [13], it was illustrated how reasoning capability of LLMs can be enhanced highly by using Chain-of-Thought, which we will be using for predicting the truthfulness value of the tweets. The following is the example of the system prompt that was used for generating the CoT Reasoning -

**System Prompt:** *You are an expert journalist. You do not hallucinate. Your views are not aligned to any political party or propaganda. You are unbiased, neutral and faithful to the Text and News Summary provided. Ignore any bias mentioned in the news. You will be provided with a Text and a News Summary. You need to identify if a Text is truthful to the News Summary. Explain step by step why you think the Text is truthful and faithful to the News Summary or the Text is not truthful or faithful to the News Summary.*

**Text:** *Processed Tweet*

**News Summary:** *Summary of Top-5 retrived articles*

---

**Response:** *Chain-of-Thought Reasoning of whether the given Text is faithful/truthful to the News Summary or not*

**3.3.4. Truthfulness Prediction.** Finally, once we have the CoT Reasoning, we use the processed tweet, and the summarized article to predict the Truthfulness of the tweet with respect to the summary of the retrieved top-5 news articles as depicted in Figure 2. We observe that our ***vanilla RaLLM***, a zero-shot CoT based technique, produced a very high Precision result with low Recall score. Even though for our usecase, high Precision is desirable, very low recall is not desirable. We believe with proper finetuning of Large Language Model for predicting the truthfulness value of tweet, we can achieve better result.

## 3.4. Ensemble Model

From our statistical modeling, we had found the LightGBM model to be the best performing one, outperforming the baseline by the Truthseeker team. To improve our overall performance, we used ensemble of both

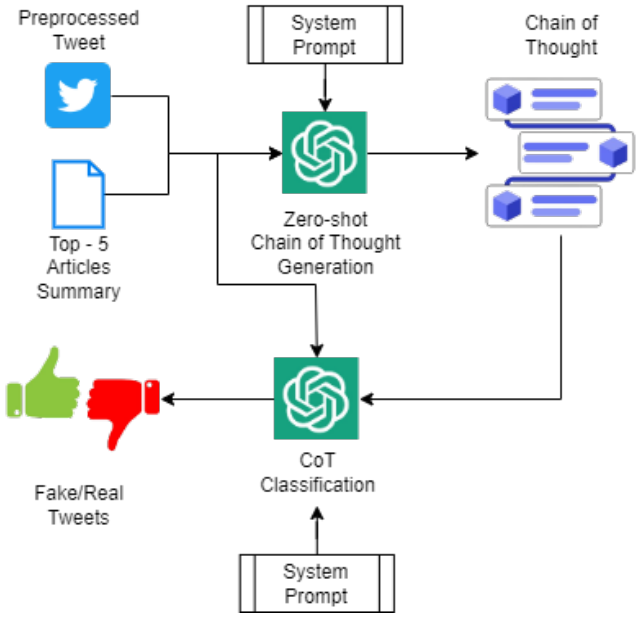


Figure 2. Chain-of-Thought Classification Framework

the models - the LightGBM and the vanilla RaLLM. With emperical evidence, we finally converged at the ensemble weightage of 80% to the prediction probability score of the LightGBM Model and a 20% weightage to the prediction of our vaniall RaLLM model. The final evaluation metrics are presented in Table 4

TABLE 4. FINAL MODEL EVALUATION: (↑) HIGHER IS BETTER (↓) LOWER IS BETTER

| Model | TPR(↑) | FPR(↓) | Precision | Recall | F1 | Accuracy |
|---|---|---|---|---|---|---|
| Random Forest | **0.70** | 0.30 | 0.70 | 0.70 | **0.70** | 0.70 |
| LightGBM | 0.69 | 0.28 | 0.71 | 0.69 | **0.70** | 0.70 |
| RALLM | 0.32 | **0.07** | 0.74 | 0.30 | 0.43 | 0.65 |
| Ensemble (LightGBM + RALLM) | 0.63 | 0.14 | **0.76** | 0.63 | **0.70** | **0.76** |

## 4. Model Evaluation

We have already established that we need low False Positives and high Precision. We see that our ensemble model of LightGBM and RaLLM provides with the highest Precision and Accuracy of all the models. There is a drop in Recall, but still the ensemble model performs as good as the Random Forest and vanilla LightGBM model in terms of the F1 score. Coming to the False Positive Rate, we see that vanilla RaLLM performs extremely good, however it loses out on Recall. Our Ensemble model provides a great trade-off by outperforming Random Forest and LightGBM with a big margin and not losing out too much on Recall. Finally, for our usecase, our Ensemble model performs the best by outperforming every other model in Precision, Accuracy and F1 score while having a commendable trade-off on

the FPR and Recall. Our Vanilla RaLLM works realtime when a tweet is posted and can help in flagging it as soon as it is posted, hence preventing the fast propagation and leading to much reduced false tweets. With the ensemble model, we achieve the state of the art performance on detecting fake tweets.

## 5. Limitations and Future Works

- Currently our model is heavily dependent on ChatGPT which is not easy to scale from an academic point of view. We only analyzed 150 tweets as our test set to reach at our evaluation metrics. Hence, we need to implement open-sourced Large Language Models in future and work to get similar accuracy as we have now.
- The faithfulness of a tweet is directly proportional to the quality of retrieved articles. Hence, we need to research on ways of retrieving efficient and better quality of articles which are congruent to the topic of the keywords.
- Chain-of-Thought and Prompt Engineering does improve the output of LLMs, however for best result we need to finetune our model for the downstream tasks [11]. As the immediate next step, we will finetune a open-source LLM to generate the faithfulness label by comparing the summary and the original tweet.

## 6. Conclusion

The extreme consumption of social media content has given rise to the propagation of fake news faster and smoother. Twitter(now X) is one of the most influential social media used for World News, Sports, Technology, Entertainment and many more. It has been studied that the number of fake tweets have shooted up in the last couple of years and the proportion of fake tweets from verified users have also sky rocketed. Statistical techniques such as Random Forest, AdaBoost, LightGBM, etc. focuses on the numerical features such as friends_count, influence, cred, which are more user-centric than tweet-centric. Fake tweet delivers False information in the disguise of a genuine tweet. The best way to identify a truthfulness of a tweet is by fact-checking it with the current true news and legitimate sources. However, doing this manually is almost impossible. Our method proposes a novel technique of Retrieval Augmented Large Language Model which retrieves the top legitimate articles relevant to a tweet and decides whether the tweet is truthful to the news or not. We showcased that an ensemble of a statistical(LightGBM) and RaLLM outperformed every other model for this task. We aim to work on the finetuning of Large Language Model and aim to build a self sufficient RaLLM system to detect the truthfulness of a tweet.

## References


[1] S. Helmstetter and H. Paulheim, "Collecting a large scale dataset for classifying fake news tweets using weak supervision", *Future Internet*, vol. 13, no. 5, p. 114, 2021.

[2] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of economic perspectives*, vol. 31, no. 2, pp. 211–236, 2017.

[3] S. Helmstetter and H. Paulheim, "Weakly supervised learning for fake news detection on twitter," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (ASONAM). IEEE, 2018, pp. 274–277.

[4] S. Dadkhah, X. Zhang, A. G. Weismann, A. Firouzi and A. A. Ghorbani, "The Largest Social Media Ground-Truth Dataset for Real/Fake Content: TruthSeeker," in *IEEE Transactions on Computational Social Systems*, 99. 1-15, Oct. 2023. [Truthseeker Dataset]

[5] T. Murayama, "Dataset of fake news detection and fact verification: A survey," *arXiv preprint arXiv:2111.03299,* 2021.

[6] A. Vlachos and S. Riedel, "Fact checking: Task definition and dataset construction," in *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, 2014, pp. 18–22.

[7] Lobato et al. "Fact-checking on Twitter: An analysis of the hashtag #StopBulos". *Revista Interamericana De Psicología/Interamerican Journal of Psychology*, 55(2), e1371.

[8] Veselovsky, Veniamin, et al. "Generating Faithful Synthetic Data with Large Language Models: A Case Study in Computational Social Science." *arXiv preprint* arXiv:2305.15041 (2023).

[9] Lanham et al. "Measuring Faithfulness in Chain-of-Thought Reasoning", *arXiv preprint* arXiv:2307.13702, 2023

[10] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel et al., "Emergent Abilities of Large Language Models", arXiv:2206.07682, 2022

[11] Kai Lv, Yuqing Yang, Tengxiao Liu, Qinghui Gao, Qipeng Guo, Xipeng Qiu, "Full Parameter Fine-tuning for Large Language Models with Limited Resources", arXiv:2306.09782, 2023

[12] Edward J. Hu, Yelong Shen, Phillip Wallis, "LoRA: Low-Rank Adaptation of Large Language Models", arXiv:2106.09685 (2021)

[13] Jason Wei, Xuezhi Wang, Dale Schuurmans et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models", arXiv:2201.11903, 2023