Fighting Lies with Intelligence: Using Large Language Models and Chain of Thoughts Technique to Combat Fake News

Waleed Kareem(✉) and Noorhan Abbas

University of Leeds, Leeds, UK
od21wk@leeds.ac.uk

Abstract. The proliferation of fake news in the digital age presents a substantial challenge, outpacing the capabilities of conventional fact-checking methods. To address this, we introduce a pioneering strategy that utilizes fine-tuned Large Language Models (LLMs) for discerning fake news through the generation of logical reasoning that validates or critiques news headlines. This strategy seamlessly merges the predictive prowess of LLMs with the requisite for coherent explanations, facilitating not only the detection of fake news but also offering transparent, reasoned justifications for each classification. Leveraging the inherent "Chain of Thought" (CoT) reasoning and model distillation processes of pre-trained LLMs, our approach enhances detection accuracy while rendering the models' complex decisions accessible to human understanding. This research signifies a groundbreaking contribution, extending beyond mere methodological progress by presenting an open-source dataset fortified with CoT annotations, establishing a new benchmark for fake news detection. This dataset, consisting of a diverse mixture of human-annotated news and those generated under human-guided contexts using the OpenAI GPT 3.5 model, promises to be a valuable resource for future scholarly endeavours in the field. By optimizing two distinct LLMs (FLAN-T5 and Llama-2), our methodology demonstrates unprecedented efficacy, surpassing the existing state-of-the-art results by 11.9% and elevating the overall performance of LLMs in fake news detection.

Keywords: Fake News Detection · LLM · COT

1 Introduction and Background

In the digital age, the rapid spread of fake news seriously threatens democracy, public health, and societal trust, potentially altering electoral outcomes and impeding crisis responses, including during the COVID-19 pandemic [3, 4]. This research endeavors to mitigate the far-reaching impacts of misinformation by harnessing the latest advances in Natural Language Processing (NLP) for more effective fake news detection. Utilizing LLMs such as FLAN-t5 and Llama-2, trained extensively on substantial text datasets, this study seeks to discern linguistic patterns and nuances integral to detecting misleading narratives. Incorporating the CoT framework, we aim to foster transparency and

M. Bramer and F. Stahl (Eds.): SGAI 2023, LNAI 14381, pp. 253–258, 2023.
https://doi.org/10.1007/978-3-031-47994-6_24

user trust in fake news detection by mimicking structured human reasoning processes, thus addressing vital issues in AI interpretability [1]. While initial mitigation efforts had limitations, modern techniques employing sentiment analysis and deep learning models promise superior accuracy in combatting evolving fake news narratives. Notably, a deeper understanding of linguistic nuances and visual cues has been highlighted as pivotal in enhancing detection mechanisms, despite the persistent challenges posed by sophisticated misinformation generators [5, 6, 8, 9]. This study underscores a continual need for innovation in the sector to keep pace with the changing tactics of fake news creators.

2 Methodology

This study adopted a rigorous data collection and processing methodology, involving data gathering from online sources, data generation using a pre-trained LLM, and then the refinement of these data using supplementary datasets. We detail the various stages of this process in the sections below.

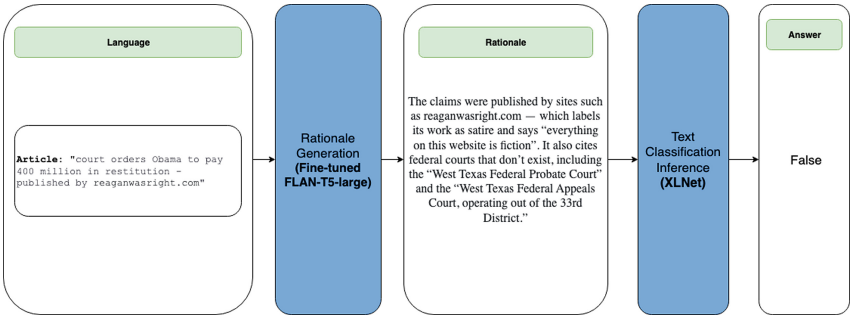

Fig. 1. An overview of the fact checking process

The flowchart in Fig. 1 maps out the fact checking process. First, we generate the rationale using the news headlines as an input to the fine-tuned LLM, then the generated text is used as an input to the final classification stage. The figure illustrates the intricate processes of data augmentation, model distillation, and text classification, which assist in visually grasping the complexity and coherence of our methodology. The initial dataset was procured from the PolitiFact.com website. This dataset encapsulates a variety of political news, including the headline, the source of the headline, and an accompanying article that provides a substantiation for the veracity of the news—categorised as true, mostly true, half true, or false.

2.1 Data Augmentation Using a LLM

The raw data collected from Politifact was augmented utilizing the OpenAI GPT-3.5 LLM, generating CoT justifications which are grounded in the headlines and details of each article. The following carefully crafted query was utilised for this purpose:

```
query = f"""
You are a political fact checking auditor, your job is to provide
logical reasoning for why a certain news article is either true,
half-true, false, flip-flop, pants-fire or partially true. Your
reasoning should be a paragraph of a minimum of 70 words and a
maximum of 150 words
Statement: {claim}
Conclusion: {label}
Background: {context}
Task: Write a step-by-step logical summary in a paragraph to jus-
tify the conclusion based on the statement ONLY. The background
section is only provided to help with the conclusion, ignore any
irrelevant information in this section. Do NOT refer to anything
in the context when building the summary. Start the justifica-
tion right away without quoting the statement or conclusion again
logical summary paragraph (no bullet points):
"""
```

The query has three variables: the claim, which represents the news headline; the label determined by the human auditor, which is one of 6 labels (true, false, mostly-true, barely-true, half-true and pants-on-fire) and the context, which represents the article that explains the label in detail. This process is akin to a form of 'model distillation', where we generate new data (the CoT justifications) that are highly informative and provide additional layers of depth to the dataset. It also helps to minimise the reliance on data types that could introduce unwanted noise or bias into the subsequent analysis.

In essence, model distillation is a form of transfer learning that, in our case, aims to compress the knowledge from a larger, more complex model (the teacher model - GPT-3) into a smaller, more manageable model (the student model). The distillation process occurs during the generation of the CoT information. Here, the LLM processes the contextual information from the news articles, and then provides detailed, logically coherent justification for the veracity of the news, thus creating a distilled knowledge representation. Through this distillation process, we created an enriched dataset consisting of 8597 CoT justifications [7]. These effectively serve as distilled knowledge representations that capture the complex reasoning process of the LLM, offering a deeper, more nuanced understanding of the news articles. We elaborate further on this process and its implications in the next section. The dataset was then supplemented by scientific misconception and facts to achieve a balance between the true and fake records.

2.2 Fine-Tuning LLM and Developing the Classification Model

Once the dataset was refined and balanced, we employed it to fine-tune smaller LLMs. We selected two distinct LLM architectures: FLAN-T5 LLM by Google and Llama-2 LLM by Meta, owing to their high performance in language tasks, noted by Llama-2's top rank on the Hugging Face Open LLM Leaderboard at the time of this study[1]. FLAN-T5 stands out for its adaptable architecture, facilitating ease in tuning across various NLP tasks, while Llama-2 excels in reasoning and handling extended contexts,

[1] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard.

essential for generating news headline justifications. Additionally, FLAN-T5's previous fine-tuning on the CoT dataset [10] augments its ability to craft logical justifications. For each architecture, we experimented with different model sizes in terms number of parameters. Initially, we fine-tuned T5-large (738M), x-large (3B), and T5-xxl (11B) using the Hugging Face 'Transformer' library and PyTorch framework. This was followed by fine-tuning Llama-2-7b (7B) and Llama-2-13b (13B). A pivotal aspect of our experimentation was tasking these student models with generating CoT (provide definition or context) justifications based on news titles and headlines. The objective was to endow the language model with the capability to reason about and justify the veracity of a news headline, enhancing the sophistication of our approach. By leveraging model distillation, we transferred the reasoning and justification abilities of the GPT-3 model (176B) to a simpler classification model. This procedure enables the capture of essential features learned by the language model for detecting fake news. Consequently, our distilled model serves as a practical, deployable tool that can swiftly and accurately classify news headlines based on their veracity.

The last stage of our process involved refining a model to classify the outputs derived from the LLM in the previous stage. For this task, we utilised the RoBERTa model, a variant of the BERT model, but also incorporated the XLNet model to test its efficacy.

3 LLM Model Evaluation

Table 1. Evaluation of fine-tuning different open source LLMs

Model Name	CE loss	GPUs	Training time
Llama-2-13b	5.37	4 Nvidia A100	~ 2.25 h
Llama-2-7b	**5.74**	**4 Nvidia A100**	**~ 1.12 h**
FLAN-T5-large	7.93	4 Nvidia V100	~ 1 h
FLAN-T5-xlarge	6.23	8 Nvidia V100	~ 2 h
FLAN-T5-xxlarge	6.01	8 Nvidia V100	~ 5 h

The evaluation of the fine-tuned text generation models focused on accuracy and loss metrics to gauge correct predictions and error degree in justifications, respectively. Initially utilising the Cohere language model, we achieved a 60% accuracy, a 6% increment from the default setup, and decreased the loss to 1.9%. Enhancements, including adding news source and date as a context in the prompt, further increased accuracy marginally to 62.7% and reduced loss to 1.6%. Exploring the performance across various sizes of FLAN-T5 and Llama-2 models revealed diminishing returns in benefits beyond certain model sizes, as evident from stagnant performance between larger versions of both models. The experiment, constrained by GPU capacity, highlighted a plateau in performance enhancement despite increasing the model size, with a notable reduction in generation loss and hallucinations and improved knowledge retrieval in certain stages of fine-tuning. The cross-entropy (CE) loss is detailed in Table 1.

When fine-tuning Llama-2 models, we used an efficient technique for fine-tuning quantized LLMs called Quantized Low Rank Adapters (QLoRA) [11]. Using QLoRA, it was possible to fine-tune Llama-2 models on 4 GPUs for a maximum of 2.5 h. Below is an example FLAN-T5 and Llama-2 models responses compared to the same base model before fine-tuning. The highlighted part shows the model's conclusion, knowledge recollection and CoT ability:

Statement: It will only cost one million dollars to build the wall between US and Mexico.- a statement by Facebook posts on April 22, 2020.
Llama-2 fine-tuned model: The statement that it will only cost one million dollars to build the wall between the US and Mexico is false. This is because the cost of the wall proposed by President Donald Trump was estimated to be between $21.6 billion and $70 billion, depending on the design and materials used. Trump himself asked for $5.7 billion for a 200-mile segment of the wall...
Llama-2 base model response: The statement is false. The cost of building the wall between US and Mexico is estimated to be around $21.6 billion
FLAN-T5 fine-tuned model response: The statement that it will only cost one million dollars to build the wall between US and Mexico is false. According to the U.S. Department of Homeland Security, the estimated cost of building the wall is $1.6 billion.

Analysis: the Llama-2 fine-tuned model's adeptness in forming coherent and credible narratives by recollecting external factual knowledge, particularly in assessing the underestimated cost of building a US-Mexico wall. In comparison, the FLAN-T5 model, although reaching a correct conclusion, displays a tendency towards speculative reasoning and offers less logical consistency.

4 Classification Model Evaluation

We evaluated our classification model on the well-known LIAR dataset. The LIAR dataset comprises news headlines, which we harnessed to generate justification using the fine-tuned LLM. The model was then tested using the LIAR test dataset2. Our evaluation results 39.25%, 77.17%, 39.25% and 21.38% for the test accuracy, precision, recall and F1 score respectively. It is worth mentioning that our model surpassed the performance of the current top-ranked model on paperswithcode.com, which held a test accuracy of 27.4% on LIAR dataset[2]. The initial model's accuracy of 39.25% improved to 84.26% when classifications were reduced to true or false, also boosting precision to 85.33% and the F1 score to 77.41%. However, further optimization is needed to enhance recall and the F1 score.

[2] https://paperswithcode.com/sota/fake-news-detection-on-liar.

5 Conclusion

This study ventures into the scarcely explored domain of combining NLP, machine learning, LLMs, and cognitive psychology to enhance fake news detection, particularly emphasizing the nuanced integration of fine-tuned LLMs with a CoT approach to foster transparency and improve human intelligibility in AI predictions. The introduced method is innovative in its dual focus on precise detection and transparent reasoning, aiding not only in identifying fake news but also in logically justifying or refuting headlines, thus proposing a comprehensive solution to tackle misinformation. This significant advancement paves the way for future research to potentially revolutionize society's approach to combating fake news, nurturing a more informed and discerning global populace.

References

1. Wei, J., et al.: Chain of thought prompting elicit reasoning in large language models. (2022). arXiv:2201.11903
2. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: Proceedings of the NIPS 2014 Deep Learning and Representation Learning Workshop, (2014). Accessed 09 July 09 2023. https://arxiv.org/abs/1503.02531
3. Cantarella, M., Fraccaroli, N., Volpe, R.: Does fake news affect voting behaviour?, Research Policy. North-Holland. (2022). Accessed 17 July 2023). https://www.sciencedirect.com/science/article/pii/S0048733322001494
4. Kim, H.K., Tandoc, E.C.J.: Consequences of online misinformation on covid-19: Two potential pathways and disparity by eHealth Literacy, Frontiers. Frontiers. (2022). Accessed: 17 July2023. https://www.frontiersin.org/articles/10.3389/fpsyg.2022.783909/full
5. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: a data mining perspective. ACM SIGKDD Explor. Newsl **19**(1), 22–36 (2017)
6. Conroy, N.J., Rubin, V.L., Chen, Y.: Automatic deception detection: methods for finding fake news. Proc. Assoc. Inf. Sci. Technol. **52**(1), 1–4 (2015)
7. Dataset. https://huggingface.co/datasets/od21wk/political_news_justifications
8. Wang, Y., et al.: EANN: Event adversarial neural networks for multi-modal fake news detection. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery, pp. 849–857 (2018)
9. Gupta, A., Lamba, H., Kumaraguru, P., Joshi, A.: Faking sandy: characterizing and identifying fake images on Twitter during hurricane sandy. In: Proceedings of the 22nd International Conference on World Wide Web Companion, pp. 729–736 (2013)
10. Chung, H. W., et. al.: Google Scaling Instruction-Finetuned Language Models (2022)
11. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: QLORA: Efficient Finetuning of Quantized LLMs. University of Washington. Email: {dettmers, artidoro, ahai, lsz}@cs.washington.edu (2023)