# Final Project Proposal
## Scaling Up Misinformation Detection with LLMs


## Team Members
**Zou(Zoey) Yang**
**Seung Hee(Cassie) Lee**


### Introduction
Large language models (LLMs) have shown proficiency in a range of language preprocessing tasks, including classification and sentence generation. These models could serve as valuable tools for identifying misinformation on social media platforms. Given the escalating issue of misinformation, which poses a significant threat to public health, addressing this problem is paramount.

Nonetheless, detecting fake news represents a formidable challenge in the quest to combat misinformation on social media. Traditional fact-checking methods, which rely on human expert validation, are costly and difficult to scale. LLMs, through their capacity for complex language processing tasks, including classification and sentence generation, have demonstrated potential as tools for automated misinformation detection. These models, built on sophisticated algorithms and expansive datasets, are capable of parsing and interpreting vast amounts of textual information, a capability critical in identifying subtle nuances of misinformation.

Through fine-tuning LLMs, we improved the accuracy slightly higher than the pre-fine-tuning one and the empirical results from Wang's paper (2017). Among all the LLMs we trained, Bert has the highest accuracy score around 0.35 compared to Wang's best result (0.27) and pre-fine-tuning accuracy 0.28. However, the accuracy from fine-tuning Llama2 is slightly better than random guess (0.2).

### Methodology

### Datasets
The dataset we used is from a benchmark dataset for fake news detection (Wang, 2017). Below we provide the dataset statistics.

| Dataset Statistics | |
|---|---|
| Training set size | 10,269 |
| Validation set size | 1,284 |

| | |
|---|---|
| Testing set size | 1,283 |
| Avg. statement length (tokens) | 17.9 |
| Top-3 Speaker Affiliations | |
| Democrats | 4,150 |
| Republicans | 5,687 |
| None (e.g., FB posts) | 2,185 |

*Table 1: The LIAR dataset statistics*

In our study, we utilize a dataset designed for fact-checking statements made by public figures, comprising attributes such as labels, statements, subjects, speakers, job titles, state information, party affiliations, and contextual data. This dataset classifies statements into six categories: 'barely-true', 'pants-fire', 'false', 'half-true', 'true', and 'mostly-true', with 'pants-fire' indicating blatant falsehoods. These labels are approximately distributed in the training set.

## Data Preprocessing

To enhance our analysis, we preprocess this data by amalgamating the headline, subject, speaker, and speaker title into a single field, a step aimed at providing a comprehensive context for each statement. This preprocessing involves text normalization, removal of irrelevant characters, and tokenization, crucial for preparing the dataset for machine learning applications. Notably, we merge the 'pants-fire' label with 'false' to streamline the classification process, maintaining data integrity while simplifying model analysis. This approach provides a holistic view of each statement, facilitating a more nuanced understanding of public discourse veracity. Below we provide a sample for the sentences we use in the training.

*'Says the Annies List political group supports third-trimester abortions on demand. Subject: abortion. Speaker: dwayne-bohac. Speaker title: State representative.'*

## Models

In our research, we primarily focus on leveraging open-source models, specifically BERT, GPT-2, and LLaMA2, due to their widespread recognition and proven efficacy in natural language processing tasks. These models were chosen for their distinct architectures and capabilities: BERT (Bidirectional Encoder Representations from Transformers) excels in understanding the context of a word in a sentence, GPT-2 (Generative Pre-trained Transformer 2) is renowned for its text generation proficiency, and LLaMA2 is noted for its efficiency in language model adaptability. We employ these models in a sequence classification task, a critical component of our methodology aimed at categorizing textual data into predefined classes. This

task is particularly relevant for analyzing the veracity of statements, as it involves interpreting the semantic meaning of text sequences to determine their classification, a crucial step in identifying misinformation.

In our study, we employed sequence classification tasks for BERT, GPT-2 and LlaMa2. This approach leverages the respective strengths of these models in processing and categorizing textual information. Specifically, we use the text classifier method for LlaMa 2 recommended by Huggingface. We applied a prompt-based technique, as demonstrated in the following example, to facilitate the classification process:

*"### Human: The statement is to be categorized into one of these labels: true, false, half-true, mostly-true, barely-true. \n Headline: 'The Annies List political group supports third-trimester abortions on demand.' Subject: Abortion. Speaker: Dwayne Bohac. Speaker's Title: State Representative.\n\n### Assistant: [Label]."*

In this format, the input is structured to include a brief introduction, followed by the headline to be classified, the relevant subject, and information about the speaker. This structured input is then processed by the models, which are tasked with assigning an appropriate truthfulness label, represented by the placeholder "[Label]" in the prompt. This method ensures a standardized and contextual approach to classifying statements, allowing for a more nuanced and accurate analysis by the models.

## Results

For each model, we used pytorch for the implementation. Each tokenized by the models' tokenizations. In BERT and GPT-2, we set the epochs to 4 and maximum length to 256. The total number for our input tokens of the training set is 12405.

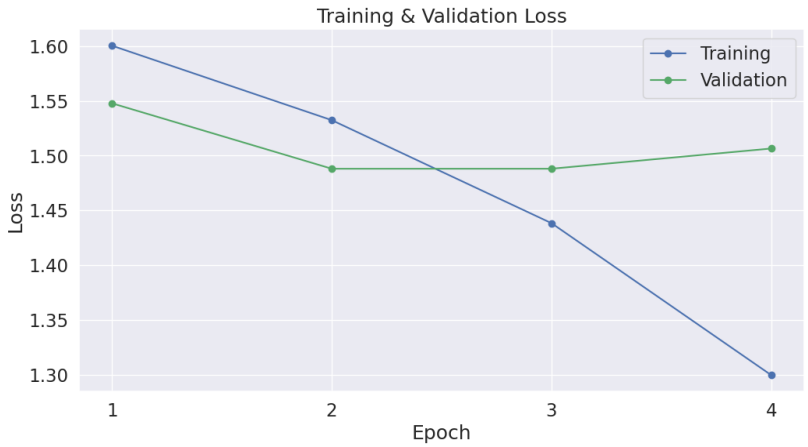


*Fig 1. Bert Large Training & Validation Loss Graph*

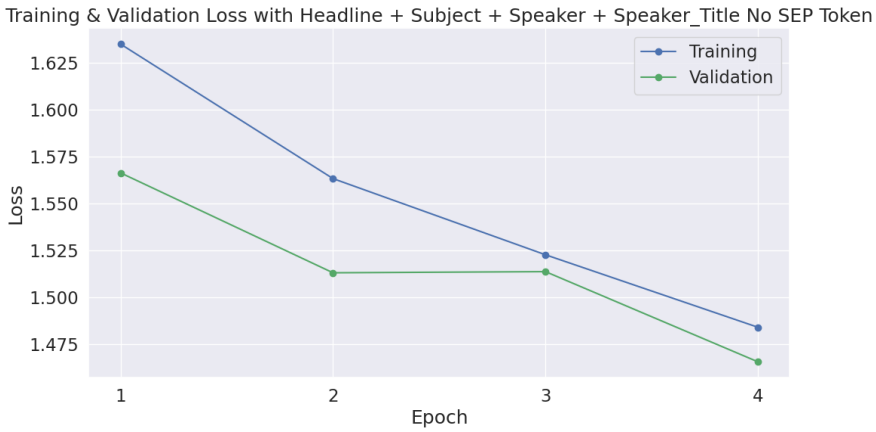


*Fig 2. GPT2 Training & Validation Loss Graph*

The above two graphs represent training and validation loss from Bert and GPT2 models over four epochs respectively. For Bert's model, the training loss shows consistent decline, indicating the model is steadily fitting the training data better. Though, validation loss seems to plateau or slightly increase between epoch 3 and 4 that there might be overfitting, or reached capacity of learning from the data. On the other hand, GPT2's graph both training and validation loss are decreasing so that model is fitting the data well and not overfitting. However, validation loss is decreasing slower which means that there could be room for further optimization.

Below we presented the results for all three models.

| **Results** | **Baseline (Wang, 2017)** | **Baseline (pre-fine tuning)** | **After-fine tuning** | | |
|---|---|---|---|---|---|
| **Models** | CNNs | Llama2 | Bert-large | Llama2 7B | GPT-2 |
| **Valid** | 0.247 | 0.0 | 0.35 | 0.203 | 0.37 |
| **Test** | 0.274 | | 0.34 | | 0.23 |

*Table 2: The evaluation accuracy on the LIAR dataset using different LLMs*

In our investigation, Table 2 meticulously outlines the empirical results, benchmarking the performance of our fine-tuned models against two primary baselines: the findings from Wang's 2017 study and the performance of the pre-fine-tuned LLaMA2 model. The results from this comprehensive comparison indicate that our fine-tuned models surpass both Wang's model and the pre-fine-tuned LLaMA2 in terms of efficacy, exhibiting an approximate 10% improvement in

accuracy following fine-tuning. Remarkably, among the evaluated models, BERT and GPT-2 show outstanding performance, BERT achieved an accuracy rate of 0.35 while GPT2 achieved 0.37 on the held-out validation set. However, it is noteworthy that the accuracy of the fine-tuned LLaMA2 model marginally exceeds that of a random guess, at 0.2. We hypothesize that this outcome may be attributed to LLaMA2 being trained on a substantially broader context compared to BERT and GPT-2, rendering our training set of 10,000 examples less impactful in enhancing its performance. This outcome particularly underscores the robustness and efficacy of BERT in our application, suggesting its strong potential as a leading model in tasks involving natural language processing.

## Conclusion

Our study demonstrates that fine-tuning Large Language Models (LLMs) enhances their efficiency in fact-checking tasks, as evidenced by the improved accuracy in classification tests. The application of fine-tuning techniques to models like BERT, and GPT-2 has shown that these models can be more effectively tailored to specific tasks, yielding better performance than their pre-fine-tuned counterparts. However, performance in LlaMa2 is not good at improving the performance. Looking ahead, we propose incorporating news domain and domain rating factors into our datasets, aligning with recent research that underscores the importance of domain context in identifying fake news. This approach is anticipated to further refine the training process, thereby augmenting the model's ability to discern and classify misinformation with greater accuracy. We want to adjust the LlaMA2 to predict the expected labels rather than random words.

## References


1. William Yang Wang. 2017. “Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.
2. Hugging Face. (n.d.). Llama2. Retrieved December 22, 2023, from *https://huggingface.co/docs/transformers/main/model_doc/llama2*