Chain of Debate: A Multi-Agent Framework for Misinformation Detection

Kovidh Jain
Virginia Tech
Blacksburg, VA, USA
kovidh13@vt.edu

Liam May
Virginia Tech
Blacksburg, VA, USA
liam23@vt.edu

Omer Abbas
Virginia Tech
Blacksburg, VA, USA
omer18@vt.edu

ABSTRACT

The deceptive confidence of LLM's, even when generating falsehoods, presents a dangerous problem, especially considering the recent explosion in use of Chat-bots. Our team, interested in the domain AI in news, investigated enhancing the credibility of the average LLM response through chain of debate. The idea is to initialize 2 opposing models that argue each other and a judge model that issues the final verdict for the most supported claim. We used Kaggle's Fake News Detection Dataset which contains various articles on controversial topics and a labels (true/false) of whether the article is fake. Our team found an improvement of .93% accuracy, 2.3% average precision, .9% average recall, .5% average F1, and a net improvement of 21 for the confusion matrix compared to a base model's response. However, with a large room for improvement, we noticed that simple classifiers such as Naive Bayes, Random Forest, and Logistic Regression are more suitable, all scoring above 94% on all metrics, which is difficult for an LLM without fine-tuning to keep up with.

1 INTRODUCTION

At the heart of large language models (LLM's) is the Transformer unit. Originally proposed by a Google research team, the Transformer was meant as a way to carry out language translation without using a recurrent neural network [6]. A recurrent neural network (RNN) is a graphical data structure that directs information cyclically to allow for variable-size input. The drawback of RNN's is that computations are done sequentially. By contrast, a Transformer block is designed to parallelize the work and coupled with GPUs, it has paved the way for deeper networks that can tackle the complexity of language, art, and music.

With adequate expressivity to create seemingly believable language, we have grown increasingly trusting of the responses, even though AI has no notion of true and false, nor any capacity to reflect on biases. For journalism in particular, LLM's can be used to create false and manipulative content. This, of course, has serious implications regarding public trust, news media reputation, and national security for adversaries.

LLM's are sensitive to the bias found in the training data. Our team's goal is to limit this by having a model refine its initial response by debating itself before giving a final result. This chain of debate will be structured similar to a formal debate with 2 models on opposing sides. The models make opening arguments, 2 rounds of response, and closing remarks. After which, a verdict will be issued by external judge model for the more compelling stance.

In terms of the conventional techniques used to address this issue there are 3:

Data Curation - This involves excluding false or harmful training examples in advance. The drawback of this approach is that it doesn't adapt to evolving information. Once a model has been trained, it's difficult and potentially expensive to retrain with a filtered dataset. Furthermore, the topic of what is harmful is subjective and unclear.

Reinforcement Learning with Human Feedback (RLHF) - This is a post-training step that has responses be reviewed by humans. This depends on a large number of expert annotators which makes this slow, expensive, and difficult to scale. Also, this might add human biases which takes the issue back to square one.

Content Filters - After a model has been trained, it's possible to effectively censor certain topics or responses. Though this can happen after training, it's not feasible to create an exhaustive list of every topic that should be banned, and again, the question of what to ban remains subjective.

These methods are either proactive and require expensive retraining or retroactive and suffer from subjectivity and bias.

2 MOTIVATION AND INTUITION

Our algorithm takes inspiration from ensemble learning. We concede that the correct response is unknown when a model is prompted. We then give equal importance to any perspective that could be taken. In binary classification, this involves creating 2 independent models that take opposing stances. This approach mirrors ensemble learning where the final result is found by synthesizing the responses of independent models, and it can be empirically shown that the sum of predictions is less likely to be wrong than any single prediction. We inferred that there are similar benefits for a debate-judge framework.

We also took loose inspiration from the Socratic method, an effective teaching strategy for guiding a learner towards the correct response through a series of question. Among humans, the Socratic method leads to understanding rather than memorization, and we inferred that this relationship should hold for a model with respect to the dataset it was trained on.

3 LITERATURE REVIEW

We have reviewed 6 papers that are novel approaches and sources of insight:

1. Chain of Agents: Large Language Models Collaborating on Long-Context Tasks - This paper discusses the benefits of using multiple agents sequentially to solve long-context tasks. This approach helps in reducing information loss in lengthy contexts, up to 10% as the compared to regular models the paper found. The idea of multi-agents was used here to prolong the context window, or how long the model is aware of prompt and response details and

sequentially. Our usage is to have the context be independent (different perspectives) and ping-pong between arguments. In reading this article, we realized that we can use multi-agents to not only to reduce information loss, but also interweave reasoning into an LLM for better responses [7].

2. MAD-Sherlock: Multi-Agent Debates for Out-of-Context Misinformation Detection - This paper proposes a multi-agent framework to evaluate information consistency. By simulating debates, it achieves state of the art accuracy in detecting the out of context & misleading prompts. Our team realized that there's no limitation in using this as an internal reasoning mechanism for any general input to see similar results in accuracy [3].

3. Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection - This paper dives into the potential of LLMs to provide multi-perspective rationale for certain news stories. The output can be used by journalist to scrutinize what they read or even inform other models to do a fact-checking. This suggests that LLM's have the potential to play the devil's advocate and argue for an alternative or perhaps unpopular side [2].

4. VeraCT Scan: Retrieval-Augmented Fake News Detection - This paper dives into retrieval-augmented systems (RAG's) for fake news detection. RAG is a technique which combines updated information/documents with generative models to increase accuracy for outputs. We experimented with making the debaters RAG's, but we were banned by browsers from scraping which we will talk about later. RAG's are generally more resistant to hallucinations, so they are a fitting augmentation to the debaters [5].

5. An Empirical Analysis on Large Language Models in Debate Evaluation - This paper dives into the capabilities and inherent biases of LLMs like GPT-4.0 as debate evaluators. It reveals that they can surpass humans and give swift and more correct judgments based on facts rather any emotions. We viewed this as promising evidence to use a judge model to assess the arguments made in the debate of the other models [4].

6. CrAM: Credibility-Aware Attention Modification in LLMs: The proposal of this paper is to augment RAG's with the ability to guage the credibility of their sources. In general, a response that's supported by a sources is better, but in some case where the sources aren't credible, this can cause a response rooted in a misinformed sources. CrAM handles this by using a system of weights such that credible resources are prioritize and less credibile ones are penalized in creating the final response [1].

4 PROPOSED TECHNIQUE

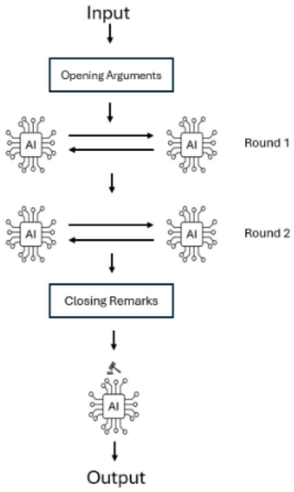


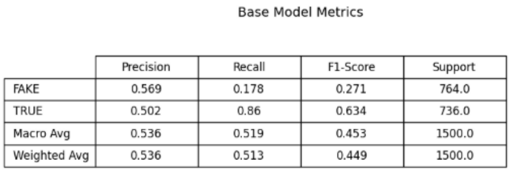
Figure 1: Debate Framework Outline

The technique we developed is called chain of debates. Given a prompt, the process begins by initializing 2 rival Falcon-7B-Instruct models with opposing context: one with with the context that a given article is true, and the other will argue that the article is untrue. The models then engage in a series of exchanges following figure 1. The structure is as follows - opening arguments, two rounds of responses, and closing remarks. As arguments are made, an independent judge model, also a Falcon-7B-Instruct and whose context is having no knowledge of the topic, silently observes the conversation. After closing remarks, the judge weighs in on the merits of each argument and selects a winner. The response is ideally the one that's most supported and technically correct. We used Kaggle's Fake News Detection Dataset (linked on our website submission) that consists of articles and labels representing whether each example is fake or true. For computational constraints, we used a randomly selected sample of 1500 examples without replacement for testing, and an additional 1500 examples were necessary to train the classifiers.

For benchmarking, we will compare chain of debate to a regular LLM as well as basic classifiers such as Naive Bayes, Random Forest, and Logistic Regression. We did so to rule out limitations in any LLM-based solution and compare our results to models that are understood to be well-suited for binary classification.

5 RESULTS

The following are the precision, recall, F1, and confusion matrix for the base model and chain of debate.



Base Model Metrics

	Precision	Recall	F1-Score	Support
FAKE	0.569	0.178	0.271	764.0
TRUE	0.502	0.86	0.634	736.0
Macro Avg	0.536	0.519	0.453	1500.0
Weighted Avg	0.536	0.513	0.449	1500.0

Figure 2: Baseline Model Metrics

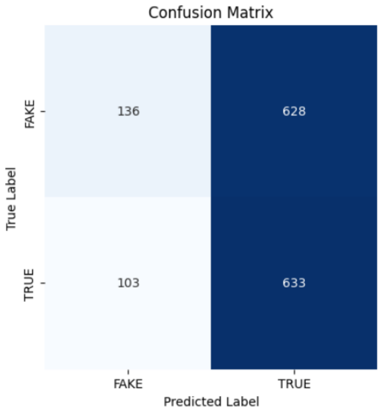
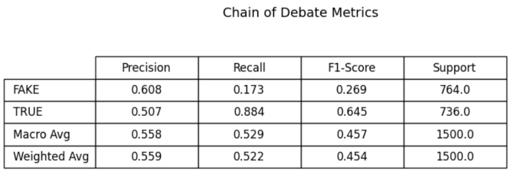


Figure 3: Baseline Model Confusion Matrix

In our testing sample, there were 764 fake examples and 736 true for a total of 1500. The base model achieved 51.26% accuracy, and a weighted average precision, recall, and F1 of 53.6%, 51.3%, and 44.9% respectively. Figure 3 shows the confusion matrix. The number of true negatives, false negatives, false positives, and true positives are found top left, bottom left, top right, and bottom right respectively. We will compare and analyze these quantities in the Evaluation section ahead. For now, consider the results of the chain of debate framework:



Chain of Debate Metrics

	Precision	Recall	F1-Score	Support
FAKE	0.608	0.173	0.269	764.0
TRUE	0.507	0.884	0.645	736.0
Macro Avg	0.558	0.529	0.457	1500.0
Weighted Avg	0.559	0.522	0.454	1500.0

Figure 4: Chain of Debate Metrics

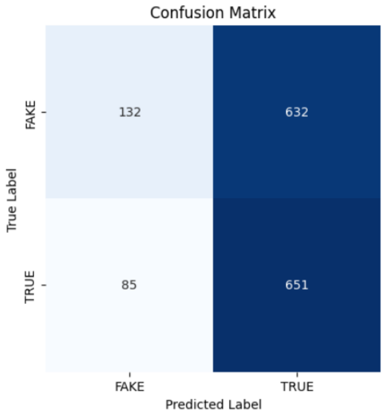


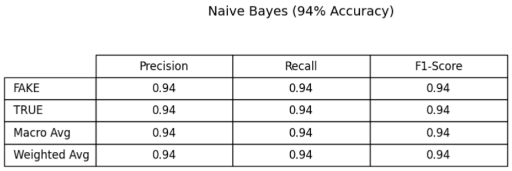
Figure 5: Chain of Debate Confusion Matrix

6 EVALUATION

We observed minor, but measurable improvement between baseline and chain of debate according to figure 4. For the evaluation metrics, chain of debate improved on accuracy and weighted (by the number of examples in each category) average precision, recall, and F1 by .93%, 2.3%, .9%, and .5% respectively. Though this is a small improvement, we were constrained by GPU's, so we had to keep the debate framework conservative, and the rounds and depth of the responses low. We saw larger gains when we used RAG's, which are LLM's augmented with source-finding capabilities, but our excessive scraping requests were detected and banned by browsers. We tabulated precision and recall for both classes since it's especially important for fake news detection. For instance, it's important that the model has a high precision for fake news which is a measure of how trustworthy such a prediction is. By contrast, recall measures how many of the positive class examples (let's say fake news) examples were correctly classified. F1 represents an average of both precision and recall, and all of these quantities can be derived from the confusion matrix.

Interestingly, in focusing on the trends for precision and recall, we noticed that both the base model and chain of debate have significantly stronger recall for true instances rather than fake instances (88% and 17% for chain of debate and 86% and 17.8% for base). What this suggests is that if an article is actually true, it's significantly more likely to be detected by an LLM than if it were fake. As it relates to precision, we can generally be marginally more trusting of a fake news prediction than a true news prediction-though the difference is not significant.

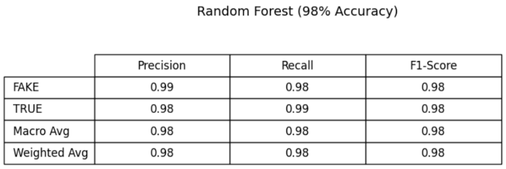
The confusion matrix enumerates all the ways a model could be right and wrong. For binary classification, this represents a 2x2 possiblity space. Chain of debate had a mixed but net improvement on the confusion matrix in figure 5. The number of true negatives (correct classifications of fake news) decreased 6; the false negatives (incorrect classifications of fake news) decreased by 15; false positives (incorrect classifications of true news) increased by 4; and true positives (correct classifications of true news) increase by 18. Overall, chain of debate had 21 more correct predictions made, taking into account the new incorrect predictions accrued.



Naive Bayes (94% Accuracy)

	Precision	Recall	F1-Score
FAKE	0.94	0.94	0.94
TRUE	0.94	0.94	0.94
Macro Avg	0.94	0.94	0.94
Weighted Avg	0.94	0.94	0.94

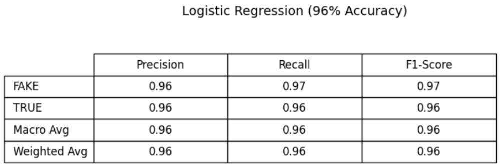
Figure 6: Naive Bayes Results



Random Forest (98% Accuracy)

	Precision	Recall	F1-Score
FAKE	0.99	0.98	0.98
TRUE	0.98	0.99	0.98
Macro Avg	0.98	0.98	0.98
Weighted Avg	0.98	0.98	0.98

Figure 7: Random Forest Results



Logistic Regression (96% Accuracy)

	Precision	Recall	F1-Score
FAKE	0.96	0.97	0.97
TRUE	0.96	0.96	0.96
Macro Avg	0.96	0.96	0.96
Weighted Avg	0.96	0.96	0.96

Figure 8: Logistic Regression Results

When comparing to Naive Bayes, Logistic Regression, and Random Forest (looking at figures 6, 7, 8), we found that these simpler models performed better, scoring above 94% on all metrics. However, it's important to note that they needed an additional 1500 samples for training, and that our chain of debate framework is a post-training optimization that can be applied to any potential LLM. Furthermore, we had to destroy the textual input using term frequency inverse document frequency (TFIDF), which is technique for converting long textual strings to numerical data by taking frequencies of the most relevant words. How the basic classifiers use this information to arrive at the final prediction is ambiguous and difficult to study. By contrast, our solution is a good technique for catching where a model goes off track since the chain of conversations remains understandable throughout the process up until the final verdict.

7 CONCLUSION AND FUTURE WORK

We set out to explore the effectiveness chain of debate in fake news detection. The results show modest but consistent improvements in accuracy, precision, recall, and F1 over a traditional single-model baseline. This approach isn't noticeably effective, especially compared to binary classifiers, but it's flexible, and there is significant room for improvement. For example, we can increase the interactions between the 2 models and the depth of the responses. Additionally, the type of model can significantly effect the performance. A constraint we had was GPU resources, so our team used the Flacon-7B-Instruct which is a conservative model for getting results in a reasonable amount of time. This can be substituted for something like GPT-4 which has more paramters and better off-the-shelf reasoning. Another simple way to improve the results would be to use a RAG for the debaters. This was the initial idea for the team, but as mentioned earlier, browsers forbid extensive scraping. This can be taken a step further by applying the CrAM model [1], so that the credibility of the source will play a factor in the final response. More importantly the debate structure helps surface logical weaknesses and improves interoperability, which makes it so that our potential journalist and readers can catch hallucinations and tangents that aren't constructive to the final response. Lastly, more experiments could be done with fine-tuning the models to the task of debate with our dataset. This is a provable strategy that has been shown to specialize an already trained LLM to a particular task.

To conclude, chain of debate is a technique that shows promise with additional research. Instead of relying on the traditional black-box approach with AI, chain of debate allows for additional transparency in the responses, and a sequence of reasoning steps that's easy to dispute and scrutinize when responses stray. With further research, chain of debate may hold the key to having an LLM supply credible responses without any pre-training or post-training efforts by model engineers.

REFERENCES

[1] Boyi Deng, Wenjie Wang, Fengbin Zhu, Qifan Wang, and Fuli Feng. 2024. CrAM: Credibility-Aware Attention Modification in LLMs for Combating Misinformation in RAG. *arXiv preprint arXiv:2406.11497* (2024).

[2] Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 20 (2024), 22105–22113.

[3] Kumud Lakara, Georgia Channing, Juil Sock, Christian Rupprecht, Philip Torr, John Collomosse, and Christian Schroeder de Witt. 2025. MAD-Sherlock: Multi-Agent Debates for Out-of-Context Misinformation Detection. In *ICLR*.

[4] Xinyi Liu, Pinxin Liu, and Hangfeng He. 2024. An Empirical Analysis on Large Language Models in Debate Evaluation. *arXiv preprint arXiv:2406.00050* (2024).

[5] Cheng Niu, Yang Guan, Yuanhao Wu, Juno Zhu, Juntong Song, Randy Zhong, Kaihua Zhu, Siliang Xu, Shizhe Diao, and Tong Zhang. 2024. VeraCT Scan: Retrieval-Augmented Fake News Detection with Justifiable Reasoning. *arXiv preprint arXiv:2406.10289* (2024).

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[7] Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan O. Arik. 2024. Chain of Agents: Large Language Models Collaborating on Long-Context Tasks. *https://arxiv.org/abs/2406.02818* (2024).